Early-Warning Prediction of Student Housing Dissatisfaction to Support Targeted University Interventions

Stefania Zourlidou[1], Kavya Hanumanthu[2]

[1, 2] Institute for Web Science and Technologies (WeST), University of Koblenz, Germany

Abstract

Student housing dissatisfaction can erode wellbeing and academic engagement, but universities often learn about serious problems only after complaints escalate. We present a leakage-aware early-warning approach that predicts serious housing dissatisfaction from a lightweight student housing survey while excluding items that directly paraphrase the target outcome. Using a University of Koblenz survey (N=158), we define an at-risk case as an overall housing experience score of 3 or lower on a 1–5 scale (51/158, 32.3%). We compare Logistic Regression, Random Forest, histogram-based gradient boosting and a compact MLP under 5-fold stratified cross-validation. Random Forest achieves the strongest discrimination (ROC-AUC = 0.803 ± 0.127), closely followed by Logistic Regression (0.792 ± 0.096). Because student support teams operate with limited outreach capacity, we move beyond headline accuracy and examine threshold policies that translate predicted probabilities into concrete workloads. At the selected operating points, Random Forest and Logistic Regression identify 38 of 51 at-risk students while flagging 64–66 students for follow-up. Permutation importance highlights expectation alignment, affordability pressure, housing instability, privacy and space limitations, and service responsiveness as the most actionable signals. The study contributes a practical blueprint for survey-based triage that links prediction, threshold selection, explainability and ethical safeguards for targeted university support.

Keywords: *early-warning systems, student housing, targeted support, explainable machine learning, threshold policies, learning analytics*

1. Introduction

Housing conditions structure students' everyday capacity to study. A room that is unaffordable, noisy, overcrowded or poorly maintained can reduce sleep quality, limit concentration and make participation in campus life more difficult. These pressures are not always visible to the university, especially when students avoid formal complaints or normalize difficult living conditions until problems become severe. Most universities therefore face a timing problem. Complaints, crisis appointments and end-of-term surveys often arrive too late for preventive support, while student services must still decide which cases deserve attention first. An early-warning signal is useful only when it is operationally realistic: it should rely on information that can be collected early, avoid label leakage from questions that directly encode the outcome, and produce decisions that match available outreach capacity.

This paper develops and evaluates such a workflow on a small institutional housing survey. We benchmark four compact tabular models, report cross-validated predictive performance, and show how decision thresholds transform risk scores into concrete support workloads. We then use model-agnostic explanations to connect risk to intervention levers, so that the output is not a black-box label but a transparent triage aid for student services. The study is guided by three research questions: **RQ1:** *To what extent can non-leaky housing survey features identify students at risk of serious housing dissatisfaction?* **RQ2:** *How do different probability thresholds change the balance between missed at-risk students and feasible outreach workload?* **RQ3:** *Which interpretable risk signals can inform support actions without reducing the model to a black-box decision tool?*

The contribution is deliberately practical and operational. First, the paper defines a leakage-aware prediction task that excludes direct satisfaction summaries from the feature set. Second, it reports not only discrimination metrics but also threshold-dependent workload indicators that a student support unit can understand. Third, it links global explanations to non-punitive support pathways, so that predictions serve as prompts for outreach rather than as labels attached to students.

2. Background and Related Work

Housing has been associated with mental health, stress and academic engagement, especially when students face affordability pressure or unstable arrangements [1,2]. These mechanisms motivate early support: if a university can identify severe dissatisfaction early, it can route students to information about contracts and rights, mediation with providers, emergency options or maintenance escalation before problems compound. Parallel work in learning analytics and early-alert systems shows that predictive models can help target support, but operational details matter as much as discrimination metrics [3-5]. In small-data institutional settings, interpretable models and clear decision policies are often preferable to complex pipelines. We therefore emphasize leakage-aware feature design, transparent thresholding and post-hoc explanations, aligning with applied machine learning guidance on cost-sensitive decision making [5,6].

This article is part of a broader line of analyses using the Koblenz student-housing survey, but it addresses a different operational question. Prior same-dataset work examined affordability and accessibility through geospatial and regression indicators [12], choice scarcity and spatial equity through mixed methods and geographically weighted regression [13], and recurring need profiles through Bernoulli latent class analysis [14]. The present paper does not reuse those analytical framings as its contribution. It instead asks whether early survey signals can be converted into a workload-aware triage policy for targeted support. Accordingly, the outcome, modelling task, validation emphasis and decision logic are distinct from those earlier studies.

Early-warning models also raise two design questions: how to communicate risk scores responsibly and how to convert them into action. In education, this conversion is often hidden behind a default threshold such as 0.5. The best operating point depends on the relative cost of missed severe dissatisfaction, unnecessary outreach, staff availability and the consequences of being flagged [6,11]. We therefore report both discrimination and workload-facing metrics, treating the threshold as a policy parameter rather than a purely technical choice.

3. Data and Problem Formulation

We use a cross-sectional student housing survey conducted at the University of Koblenz (N=158). The instrument includes demographic and study context, housing characteristics, search and moving history, decision factors, service responsiveness, perceived fit and overall housing experience. The survey was designed for institutional insight rather than high-stakes classification, which makes it a realistic test case for low-burden early warning.

The prediction task is binary classification. A respondent is labelled at-risk if the overall housing experience score is 3 or lower on a 1–5 Likert scale (51/158; 32.3%). This threshold captures clearly negative or ambivalent experiences where outreach is plausibly warranted. To avoid label leakage, we remove survey items that directly summarise satisfaction or overall experience. The remaining features describe housing conditions and potential drivers of dissatisfaction, not the target itself.

For interpretation, features are grouped into five domains: affordability and financial strain, expectation and information alignment, service and maintenance experience, physical space and privacy, and stability or search history. This grouping supports communication with non-technical stakeholders and discourages over-interpretation of isolated questionnaire items. Class imbalance is moderate, so we rely on stratified validation and threshold tuning rather than aggressive resampling, which can be unstable in small samples.

4. Methods

Feature representation and preprocessing. Numerical variables such as room size, semester number and number of moves are used as numeric inputs. Ordinal Likert items are encoded with ordered integer scores; categorical variables are one-hot encoded. Missing values are imputed within the modelling pipeline using fold-consistent rules (median for numeric variables and mode for categorical variables) to avoid information leakage across folds.

Models. We benchmark four models that can be deployed without heavy infrastructure: Logistic Regression with L2 regularization, Random Forests [7], histogram-based gradient boosting and a compact multi-layer perceptron. The goal is not to maximise marginal gains, but to identify a model class that balances discrimination, stability and interpretability for institutional use.

Validation, metrics and threshold policies. We use 5-fold stratified cross-validation with shuffling and a fixed random seed (seed = 42). Discrimination is assessed with ROC-AUC, while operational performance is assessed with precision, recall and F1. Because support services operate under capacity constraints, we tune the decision threshold within the cross-validation procedure and report deployment-

style confusion matrices that make workload explicit: how many students are flagged, how many at-risk cases are found and how many are missed. These thresholds should therefore be interpreted as planning-oriented operating points for this survey wave, not as final deployment thresholds. A live system would require prospective validation on a later survey wave before routine use.

Probability calibration. Threshold policies rely on the numerical meaning of predicted probabilities, so discrimination alone is not enough. A model may rank students well while still over- or under-estimating absolute risk. Before deployment, predicted probabilities should therefore be checked with calibration diagnostics such as reliability plots or Brier score, and recalibration methods should be considered if probabilities are poorly calibrated [10]. In this paper, calibration is treated as a deployment requirement and not as evidence already established by the present sample.

Explainability. For the best-performing model, we compute permutation importance using ROC-AUC as the scoring function. This model-agnostic ranking identifies which variables most influence predictive performance and helps map risk signals to actionable responses. Logistic Regression coefficients provide a second, simpler interpretive check; in a future deployment, local explanation methods such as LIME and SHAP could support student-specific review [8,9].

5. Results

5.1 Predictive Performance

Table 1 reports mean performance across the 5-fold stratified validation and should be read in two ways: as a comparison of model discrimination and as evidence that the classification threshold materially changes operational performance. Random Forest provides the strongest discrimination (ROC-AUC = 0.803 ± 0.127) and the best threshold-tuned F1 (0.733). Logistic Regression follows closely (ROC-AUC = 0.792 ± 0.096, tuned F1 = 0.695), which means that the final model choice need not be based on accuracy alone.

Model	ROC-AUC (mean ± sd)	**F1 @0.5 (mean)**	**F1 tuned (mean)**	**Precision tuned (mean)**	**Recall tuned (mean)**	**Threshold (mean)**
Random Forest	0.803 ± 0.127	0.635	0.733	0.708	0.802	0.40
Logistic Regression	0.792 ± 0.096	0.607	0.695	0.623	0.802	0.27
HistGB	0.767 ± 0.131	0.571	0.652	0.618	0.724	0.28
MLP	0.649 ± 0.116	0.379	0.551	0.435	0.800	0.29

Table 1. Model comparison under 5-fold stratified cross-validation. Recall reported here is the mean across folds at per-fold tuned thresholds. Pooled out-of-fold recall at fixed thresholds in Table 2 differs slightly.

The most important pattern in Table 1 is that threshold tuning improves F1 for every model compared with the default 0.5 threshold. This is expected in a support setting where the positive class is smaller than the negative class and where missed at-risk cases are operationally important. The tuned thresholds therefore determine how many students become visible to the support workflow. The table also shows a practical trade-off between marginal predictive gain and institutional transparency. Random Forest is the strongest performer, but Logistic Regression remains close enough to be a credible alternative if a university prioritises a simpler scoring rule for communication and review. By contrast, the MLP has lower ROC-AUC and F1 in this small tabular setting, so the results favour compact, interpretable or ensemble methods over a more complex neural baseline.

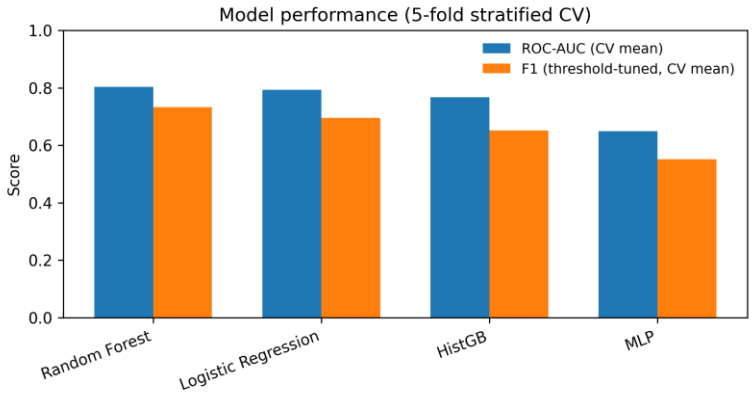

Fig. 1. Cross-validated performance summary. ROC-AUC reports discrimination, and threshold-tuned F1 reflects the per-fold optimal threshold.

5.2 Operational Thresholds and Workload

Accuracy alone does not determine whether an early-warning system is usable. The decision threshold determines how many students are routed to follow-up. Table 2 illustrates deployment-style confusion matrices at the selected thresholds (0.40 for Random Forest, 0.27 for Logistic Regression). At these settings, both models identify 38 of 51 at-risk students while flagging roughly 64–66 students in total. A support unit can move along this precision-recall trade-off by selecting a threshold that matches outreach capacity. Because Table 1 reports fold-level means at per-fold tuned thresholds whereas Table 2 reports pooled counts at fixed thresholds, the recall values differ slightly.

Model (threshold)	TN	FP	FN	TP	Flagged (TP+FP)	Missed (FN)	Precision	Recall
Random Forest (0.40)	81	26	13	38	64	13	0.594	0.745
Logistic Reg. (0.27)	79	28	13	38	66	13	0.576	0.745

Table 2. Deployment-style confusion matrices and workload metrics at fixed thresholds.

5.3 Interpretable Risk Drivers

Permutation importance for the Random Forest indicates that expectation alignment is the dominant signal, followed by affordability pressure, indicators of housing instability, privacy or space limitations and service responsiveness. This pattern points to concrete levers, including clearer listings and realistic photos, contract briefings, affordability guidance, emergency support and faster escalation for recurring maintenance issues.

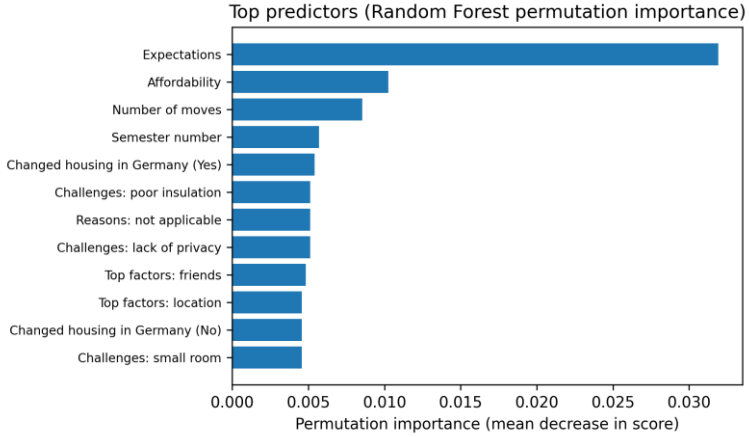

Fig. 2. Top predictors from Random Forest permutation importance (y-axis labels are survey item short names).

As a worked triage example, suppose a support unit can conduct 50 short check-ins in the next two weeks. The threshold can be calibrated to yield approximately 50 flagged students, prioritising the highest predicted risks while keeping workload feasible. Lowering the threshold would increase recall but require more outreach; raising it would reduce staff load but miss more at-risk students. This illustrates why thresholding should be treated as a governance decision rather than a fixed modelling default.

6. From Risk Scores to Responsible Outreach

The intended output of the model is not an automated decision, but a structured prompt for earlier support. Table 3 translates the modelling workflow into an institutional process that can be audited and adjusted. The table is deliberately framed as a support workflow: every step leaves room for human judgement, documentation and revision.

Table 3. Operational workflow for a leakage-aware housing early-warning process.

Step	Input	Decision point	Support-oriented output
1. Early survey	Short housing questionnaire	Exclude outcome-like satisfaction summaries	Leakage-aware feature set
2. Risk scoring	Preprocessed survey features	Apply trained model and record uncertainty	Ranked list of risk scores
3. Threshold choice	Current staff capacity and risk tolerance	Select threshold or top-k workload rule	Feasible flagged group
4. Triage review	Risk score plus global risk-driver information	Match support intensity to likely need	Information, check-in, mediation or escalation
5. Monitoring	Outreach outcomes and error patterns	Audit calibration, subgroup performance and workload	Revised threshold, survey items or workflow

6.1 Thresholds as Capacity Decisions

The same model can support different operating modes depending on the institutional situation. A lower threshold is appropriate when the priority is to miss as few severe cases as possible, while a higher threshold or top-k rule is appropriate when staff time is scarce. Table 4 describes policy choices that can be made without adding new model evidence or changing the underlying survey.

Table 4. Example threshold-policy choices for different support contexts.

Support context	Threshold logic	Priority metric	Reason
Start of semester or known housing pressure	Use a lower threshold	Recall	Avoid missing students whose problems may escalate quickly
Normal week with limited staff	Use a higher threshold or top-k list	Precision and workload	Keep outreach feasible and reviewable
Pilot deployment	Begin conservatively and manually review flagged cases	Trust and auditability	Test workflow before routine use
After additional survey waves	Recalibrate and revisit threshold	Calibration and stability	Check whether probabilities and workload remain reliable

6.2 Appropriate and Inappropriate Uses

The model should be used only to prioritise supportive contact. A flagged case should trigger an invitation to help, not an administrative label. A non-flagged case should not be interpreted as evidence that a student has no housing problem. The system should not be used to allocate scarce housing places, evaluate tenancy behaviour, penalise students or landlords, or replace normal access to support services. These restrictions are important because the target is self-reported dissatisfaction, the sample is small, and prediction errors are unavoidable.

6.3 Governance Checks Before Live Use

Before live deployment, the workflow should pass four checks. First, students should know that survey responses may be used to offer support and should understand what will not happen as a result of being flagged. Second, the model should be prospectively validated on a later survey wave, including calibration checks and subgroup error review. Third, staff should document the selected threshold, the capacity assumption behind it and the date of review. Fourth, outcomes of outreach should be evaluated qualitatively and quantitatively so that the system is judged by whether it improves support, not only by whether it predicts a survey label.

7. Discussion and Implications

The main design lesson is that prediction becomes educationally useful only when it is connected to a concrete service process. In this case, the model is best understood as a triage instrument. It helps identify students who may benefit from earlier contact, and the form of that contact remains a human support decision. The explanation profile matters because it connects risk scores to actions that the institution can take. Expectation alignment suggests the need for clearer pre-arrival information and more realistic accommodation descriptions. Affordability pressure suggests referral to financial counselling or emergency support. Maintenance and service signals suggest response-time standards, clearer reporting channels and escalation rules. In this way, the model output becomes a guide for practical support, not only a statistical result. Threshold selection is the point at which analytics becomes policy. The selected threshold reflects an institutional judgement about missed severe dissatisfaction, unnecessary outreach and available staff time. For that reason, the threshold should not be hidden inside the software pipeline. It should be documented, periodically reviewed and communicated to the team using the system [6,11].

Ethical safeguards are also methodological safeguards. Consent, data minimisation and separation between support and enforcement reduce the risk that students experience the system as surveillance. Error audits are equally important. False positives can create unnecessary concern, and false negatives leave students unsupported. Human review and transparent communication are therefore part of the validity of the intervention, not optional additions.

The study is limited by its small, single-institution and cross-sectional sample, and by a self-reported target label. The reported thresholds and confusion matrices are useful for planning, but they do not replace prospective validation. The present paper also does not estimate causal effects of outreach. It shows how risk can be identified and operationalised, but it does not test whether a specific intervention changes housing outcomes. Future work should validate the workflow across survey waves and universities, assess calibration before deployment, audit subgroup error patterns, and evaluate whether outreach improves housing conditions, wellbeing or engagement.

8. Conclusion

We presented an early-warning system that predicts serious student housing dissatisfaction from a lightweight survey and translates predictions into feasible outreach decisions through threshold policies. In this sample, Random Forest achieved the strongest discrimination. Logistic Regression remained close enough to be a plausible lower-complexity alternative when communicability, auditability and ease of recalibration matter. The broader methodological lesson is that small-data institutional settings require more than headline metrics. Models become more useful when discrimination, threshold-dependent workload and ethical safeguards are reported together. The workflow connects prediction to concrete support levers such as expectation management, affordability guidance and maintenance escalation. In this way, predictive analytics becomes a practical and auditable tool for student wellbeing, not an abstract score.

REFERENCES

[1] Evans G.W., Wells N.M., and Moch A., "Housing and mental health: A review of the evidence and a methodological and conceptual critique", Journal of Social Issues, vol. 59, no. 3, 2003, pp. 475-500. DOI: 10.1111/1540-4560.00074.

[2] Awcock H., "Edinburgh's Student Housing Crisis: The Impact of Insecure Housing on Student Wellbeing and Engagement", Student Engagement in Higher Education Journal, RAISE Network, 2025, pp. 216-233. DOI: https://doi.org/10.66561/sehej.v7i2.1256

[3] Siemens G. and Baker R.S., "Learning analytics and educational data mining: towards communication and collaboration", Proc. 2nd International Conference on Learning Analytics and Knowledge (LAK 2012), ACM, 2012, pp. 252-254. DOI: 10.1145/2330601.2330661.
[4] Jayaprakash S.M., Moody E.W., Lauria E.J.M., Regan J.R., and Baron J.D., "Early alert of academically at-risk students: an open source analytics initiative", Journal of Learning Analytics, vol. 1, no. 1, 2014, pp. 6-47. DOI: https://doi.org/10.18608/jla.2014.11.3
[5] Baker R.S. and Inventado P.S., "Educational data mining and learning analytics", in Learning Analytics, In Larusson J. and White B. (eds.), Springer, 2014, pp. 61-75. DOI: 10.1007/978-1-4614-3305-7_4.
[6] Provost F. and Fawcett T., Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media, 2013.
[7] Breiman L., "Random forests", Machine Learning, vol. 45, 2001, pp. 5-32. DOI: https://doi.org/10.1023/A:1010933404324
[8] Ribeiro M.T., Singh S., and Guestrin C., "Why should I trust you? Explaining the predictions of any classifier", Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), ACM, 2016, pp. 1135-1144. DOI: 10.1145/2939672.2939778.
[9] Lundberg S.M. and Lee S.-I., "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems 30 (NeurIPS), Curran Associates Inc., 2017, pp. 4765-4774.
[10] Niculescu-Mizil A. and Caruana R., "Predicting good probabilities with supervised learning", Proc. 22nd International Conference on Machine Learning (ICML), ACM, 2005. DOI: 10.1145/1102351.1102430.
[11] Elkan C., "The foundations of cost-sensitive learning", Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), 2001, pp. 973-978. DOI: 10.5555/1642194.1642224.
[12] Hanumanthu K., Zourlidou S., and Hopfgartner F., "Who Pays, Who Reaches? A Data-Driven Analysis of Student Housing Affordability and Accessibility", IMCL 2025 conference paper, 2025 (to appear).
[13] [Hanumanthu K. and Zourlidou S., “From Choice Scarcity to Spatial Equity: A Mixed-Methods Analysis of Student Housing in a German University City”, INTED2026 Proceedings, 20th International Technology, Education and Development Conference, Valencia, Spain, 2–4 March 2026, article 1313, IATED, 2026. DOI: 10.21125/inted.2026.1313.
[14] Hanumanthu K. and Zourlidou S., "Understanding Student Housing Needs in Higher Education: Recurring Situations, Different Support", EDULEARN 2026 conference paper, 2026 (to appear).