Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

A Higher Threshold Can Increase Error Cost Even When False Alarms Cost More: An Office-Sensor Holdout

Brandon Matthews[1] · Trevor Russell[2] · Matthew Fisher[3]

[1] Slovakia · [2] Slovenia · [3] Cyprus

Original publication: November 5, 2026 | Article 683

Abstract

Raising an occupancy threshold from 0.5 to 0.8 removed seven false alarms but added 376 missed occupied observations in a later office recording. Even when a false alarm was assigned four times the cost of a miss, the total weighted error increased. Lowering the threshold to 0.2 had the opposite tradeoff and reduced aggregate error when misses received fourfold weight. Equal-day contrasts remained uncertain across only eight test dates. We obtained these results from one frozen logistic sensor model and three fixed thresholds. The example shows why a cost preference does not, by itself, establish that a particular threshold change improves the observed decision outcome.

Public sensor-data analysis with explicitly assumed error costs. No building-control intervention or new occupancy study was conducted.

A Higher Threshold Can Increase Error Cost Even When False Alarms Cost More: An Office-Sensor Holdout

Public sensor-data analysis with explicitly assumed error costs. No building-control intervention or new occupancy study was conducted.

Abstract

Raising an occupancy threshold from 0.5 to 0.8 removed seven false alarms but added 376 missed occupied observations in a later office recording. Even when a false alarm was assigned four times the cost of a miss, the total weighted error increased. Lowering the threshold to 0.2 had the opposite tradeoff and reduced aggregate error when misses received fourfold weight. Equal-day contrasts remained uncertain across only eight test dates. We obtained these results from one frozen logistic sensor model and three fixed thresholds. The example shows why a cost preference does not, by itself, establish that a particular threshold change improves the observed decision outcome.

1. A directional intuition needs an empirical check

An occupancy model returns a score, but a building application must decide how to act on that score. A higher threshold seems attractive when false alarms are expensive, because it admits fewer positive classifications. The same change can also create missed occupied observations. Whether the trade is worthwhile depends on the number of each error that actually changes, rather than on the direction of the threshold alone.

We ask whether two fixed threshold changes improve two declared error-cost objectives when a model is transferred to a later recording. One objective assigns missed occupancy four times the weight of a false alarm. The other assigns a false alarm four times the weight of a miss. These are illustrative preferences, not costs measured from occupants, equipment, or electricity bills. The empirical object is how the fixed model's observed classifications respond to the declared choices.

Work on robust signal estimation such as Yan et al., 2021 provides relevant background for separating an estimation procedure from its downstream use. We do not implement its graph-based adaptive estimator. The present study uses logistic regression to produce occupancy probabilities and then examines the decision threshold. A reliable signal estimator and a useful classification policy are different objects; the dataset allows a direct check of the latter under a simple, recorded prediction procedure.

2. The office resource supplies an actual later period

We use Candanedo, 2016, associated with the office-sensor study in Candanedo and Feldheim, 2016. The released files contain temperature, humidity, light, carbon-dioxide measurements, a humidity ratio, and occupancy labels. The training file has 8,143 rows from February 4 to February 10, 2015. The later test file has 9,752 rows from February 11 to February 18. We verify the temporal ordering directly rather than inferring it from the word test in a filename.

Another supplied file contains 2,665 observations from an earlier recording. It is not used here. Calling all released test files future observations would misdescribe their dates. Our design uses only the strictly later file for the threshold comparison. This keeps the temporal question coherent while acknowledging that the later sample is still one additional recording of the same office, rather than a new building or a representative set of workplaces.

The later file contains 2,049 occupied and 7,703 unoccupied observations. Both labels occur throughout the supplied evaluation sample, but adjacent minute-scale observations share occupancy episodes and environmental conditions. The number of rows therefore greatly exceeds the number of independently observed settings. No new person-level labels, occupant identities, or manual adjudications are added to the source's binary occupancy field.

Building-model discussions in Foucquier et al., 2013 and Amasyali and El-Gohary, 2018 locate the statistical modeling context, while occupant-behavior questions in Hong et al., 2017 and thermal considerations in Ruellan et al., 2016 explain why the available sensor fields cannot capture every mechanism of room use. Appliance-use evidence in Firth et al., 2008 and Jones and Lomas, 2016 concerns related but distinct household processes. We do not use those studies as additional office observations or infer appliance behavior from an occupancy classification error.

3. Freeze one probability model before changing decisions

The predictor set is temperature, humidity, light, and carbon dioxide. Each is centered and scaled using its training-file mean and population standard deviation. The later file receives exactly those saved transformations. A binomial generalized linear model with a logit link is fitted on the training observations, including an intercept. The numerical fit converges within its fixed iteration cap. Its training deviance is 956.834, reported as a fitting diagnostic rather than a later-period performance guarantee.

We do not fit separate models for separate thresholds or retune coefficients after viewing the later outcomes. All three classification rules use the same frozen probability for each later observation. A score at least 0.2, 0.5, or 0.8 produces an occupied classification under its corresponding policy. This isolates the threshold choice from differences in predictors, training sample, or model architecture.

Probability calibration work in Guo et al., 2017 concerns whether reported probabilities correspond to empirical frequencies. We do not perform a separate calibration fit, so the scores should not be assumed perfectly calibrated merely because they come from a logistic link. The cost-sensitive perspective in Elkan, 2001 explains why objectives matter to classification. It does not imply that a threshold picked from an idealized probability model will improve costs on every temporally shifted sample.

Selective-prediction ideas in Geifman and El-Yaniv, 2017 concern allowing a system to withhold a decision. Our three rules always return a binary classification; there is no abstention region. This distinction prevents a threshold experiment from being misreported as a reliability gate. A policy that declines uncertain cases would require another outcome definition and an accounting of which observations no longer receive an ordinary occupancy decision.

4. Define the two error-cost scenarios explicitly

For the miss-sensitive scenario, each false positive contributes one unit and each false negative four units. For the false-alarm-sensitive scenario, those weights are reversed. Correct classifications contribute zero. The units are weighted classification errors, not dollars, energy, or a measured comfort index. The choice of four is fixed for illustration, and no cost ratio is estimated from the later labels.

We compare threshold 0.2 against 0.5 under the miss-sensitive objective. Separately, we compare 0.8 against 0.5 under the false-alarm-sensitive objective. These are two different policy questions. A negative loss difference favors the changed threshold within its stated objective. Combining their losses into one favorable statistic would merge different preferences and obscure which kind of error drives the result.

Scoring principles in Gneiting and Raftery, 2007 distinguish evaluating a probability distribution from evaluating an action taken at a threshold. ROC concepts in Fawcett, 2006 and precision-recall discussion

in Saito and Rehmsmeier, 2015 address other useful summaries across operating points. Matthews, 1975 offers another way to summarize binary classifications. None replaces the explicit weighted-error endpoint here, and we do not report a curve or a correlation coefficient as if it established the best threshold for either assumed preference.

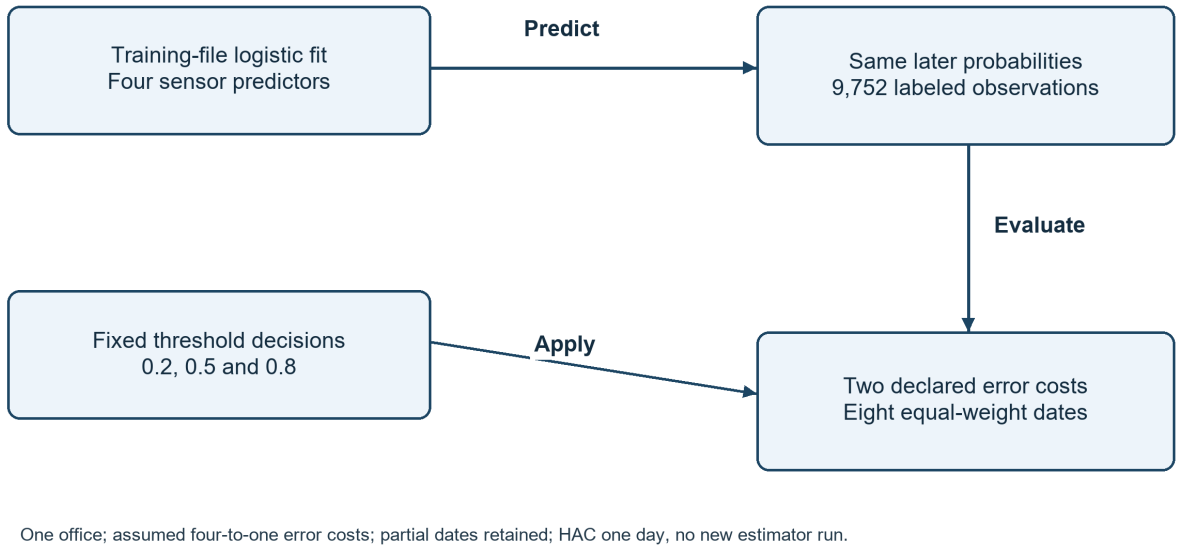

Figure 1. A common later-period probability stream enters three fixed thresholds, followed by two separately weighted error comparisons.

5. Report both interval-weighted and date-weighted evidence

The confusion counts use all 9,752 later observations. For uncertainty summaries, we first average each paired loss difference within its calendar date and then give the eight dates equal weight. Partial first and last dates are retained. Consequently, the equal-day mean can differ from the difference calculated by pooling every minute. This is a declared weighting choice, not a numerical inconsistency between the table and the aggregate error counts.

Forecast-comparison arguments in Diebold and Mariano, 1995 and West, 1996 motivate paired out-of-sample loss contrasts rather than unrelated tests of each classifier's accuracy. We fit intercept-only regressions to the daily differences and use a one-day HAC covariance based on Newey and West, 1987, with the recorded correction and a t reference using seven degrees of freedom. Eight dates provide weak support for broad temporal inference, even though the row count makes the aggregate confusion totals exact for this recording.

Table 1. Equal-day weighted-error differences under the two declared four-to-one preferences.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.0133 (0.0421)	[-0.1129, 0.0863]	8
M1: fit	R-squared = 0.0000	-	-	8
M2	Mean outcome/contrast	0.0498 (0.0355)	[-0.0342, 0.1338]	8
M2: fit	R-squared = 0.0000	-	-	8

Model labels: M1 = Threshold 0.2 minus 0.5, FN weighted four; M2 = Threshold 0.8 minus 0.5, FP weighted four.

The mean-only models have zero explanatory R-squared. Their purpose is to summarize a paired decision contrast, not explain why some dates are more difficult than others. Heteroskedasticity-only covariance in White, 1980 would not by itself represent serial dependence between daily losses. The fixed one-day HAC choice is recorded, but it cannot establish that all dependence disappears after that lag or compensate for a lack of independent office settings.

6. The lower threshold helps the miss-sensitive aggregate objective

At threshold 0.5, the model produces fifty-three false positives and 122 false negatives. Its accuracy is 0.98205. Reducing the threshold to 0.2 produces 209 false positives and thirteen false negatives, with accuracy 0.97724. Accuracy decreases because the added false alarms outnumber the removed misses. Yet the miss-sensitive objective improves because a removed miss receives four times the weight of an added false alarm.

In total, the lower threshold adds 156 false positives and removes 109 false negatives. Under the declared weights, that changes weighted error by 156 minus four times 109, a reduction of 280 units across the later file. Mean weighted error per observation falls from 0.05548 to 0.02676. These values describe the chosen scenario on the observed recording, without establishing that its assumed error weights match a particular operational application.

The equal-day contrast is -0.01331, with standard error 0.04212 and interval -0.11292 to 0.08629. The interval includes zero. Thus a favorable aggregate calculation coexists with substantial date-level uncertainty under the chosen inference. We report both rather than presenting the thousands of minute rows as thousands of independent replications or describing the changed threshold as reliably better across future dates.

The temporal-validation concerns in Bergmeir et al., 2018 and Roberts et al., 2017 remain relevant even with a genuine future holdout. The later room conditions can differ from training, and the available dates share one setting. Model-selection bias discussed by Varma and Simon, 2006 would become more serious if many thresholds and weighting rules were tried and only the best displayed. We retain the fixed three thresholds and the two specified contrasts, including their uncertain intervals.

7. The higher threshold fails its intuitive cost test

Increasing the threshold from 0.5 to 0.8 reduces false positives from fifty-three to forty-six, but raises false negatives from 122 to 498. Its accuracy becomes 0.94422. When false positives cost four units and misses one, eliminating seven false alarms saves twenty-eight units, while the 376 additional misses add 376. The net weighted error therefore increases by 348 units. The direction of the preference and the direction of a threshold change do not determine the net empirical result.

Mean false-alarm-sensitive error rises from 0.03425 to 0.06993 per observation. The equal-day difference is 0.04978, with standard error 0.03552 and interval -0.03421 to 0.13378. Its point estimate is unfavorable and its interval includes zero. We do not convert the point direction into a confirmed general penalty, but it is enough to reject the unsupported assertion that raising the threshold necessarily helps whenever false alarms receive the larger weight.

This result also does not prove that threshold 0.5 is optimal. Another threshold might offer a different balance, and a calibrated model under a stable distribution could support a more principled cost-derived rule. We have not searched for that optimum or reserved a separate threshold-development period. The

experiment addresses two specified changes from a common reference, which is a narrower and more auditable question.

8. What the data do not identify

Missing-data reasoning in Little and Rubin, 2002 and measurement-error considerations in Carroll et al., 2006 constrain the interpretation of sensor records and occupancy labels. A complete row does not ensure perfect label timing, and an observed mistake need not indicate a broken sensor. We have no independently verified list of faults or new manual occupancy review. Accordingly, neither threshold result is described as fault tolerance, and the source's room-use process is not inferred from the confusion counts alone.

Influence analysis motivated by Cook, 1977 could locate dates or observations that dominate a loss contrast. They have not been deleted because they make the date-level interval uncertain. Guidance in Wasserstein and Lazar, 2016 and the research-validity concerns in Ioannidis, 2005 favor retaining that uncertainty alongside the exact counts. Practical usefulness would require a cost model tied to an actual building decision and evaluation over enough independent conditions to assess transport beyond this office.

Documentation under Gebru et al., 2021 records file dates, the unused earlier file, training-only scaling, the frozen model, thresholds, assumed weights, and partial-date treatment. Wilkinson et al., 2016 motivates preservation of the probabilities and per-date losses. The empirical lesson is specific: a lower threshold improved one weighted objective despite lower accuracy, while a higher threshold worsened another objective despite removing false alarms. Threshold decisions need an explicit objective and observed tradeoff accounting, rather than a directional intuition alone.

The two assumed cost scenarios do not identify an actual building operator's preferences. A missed occupied interval could have different consequences at different times, and a false alarm may carry a different burden depending on the control action it triggers. Those outcomes are not observed in the sensor table. We retain constant illustrative weights to make the threshold comparison reproducible, while leaving policy selection unresolved. The frozen probabilities also come from one fitted model; changing the classifier, calibration, or training period could change which threshold has lower cost. Evaluating that broader procedure would require a new training and validation design, rather than treating the present eight dates as evidence for every occupancy-control setting.

A deployment choice would additionally require a stated action horizon. These recorded intervals do not reveal how quickly a building controller could respond or how long an occupancy error would persist after a threshold decision. Neither response time nor downstream energy use is inferred from the confusion totals.

References

1. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
2. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
3. Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2017.11.003.
4. Yan, Yi; Adel, Radwa; Kuruoglu, Ercan E.. (2021). Adaptive Normalized LMP Estimation for Graph Signal Processing. 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), 1-6. https://doi.org/10.1109/mlsp52302.2021.9596181.

5. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

6. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

7. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.

8. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

9. Firth, S.; Lomas, K.; Wright, A.; Wall, R.. (2008). Identifying trends in the use of domestic appliances from household electricity consumption measurements. Energy and Buildings. https://doi.org/10.1016/j.enbuild.2007.07.005.

10. Hong, Tianzhen; Yan, Da; D’Oca, Simona; Chen, Chien-fei. (2017). Ten questions concerning occupant behavior in buildings: The big picture. Building and Environment 114, 518-530. https://doi.org/10.1016/j.buildenv.2016.12.006.

11. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

12. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

13. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

14. Jones, Rory V.; Lomas, Kevin J.. (2016). Determinants of high electrical energy demand in UK homes: Appliance ownership and use. Energy and Buildings. https://doi.org/10.1016/j.enbuild.2016.02.020.

15. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

16. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

17. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

18. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

19. Ruellan, Marie; Park, Herie; Bennacer, Rachid. (2016). Residential building energy demand and thermal comfort: Thermal dynamics of electrical appliances and their impact. Energy and Buildings. https://doi.org/10.1016/j.enbuild.2016.07.029.

20. Amasyali, Kadir; El-Gohary, Nora M.. (2018). A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews. https://doi.org/10.1016/j.rser.2017.04.095.

21. Foucquier, Aurélie; Robert, Sylvain; Suard, Frédéric; Stéphan, Louis; Jay, Arnaud. (2013). State of the art in building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews. https://doi.org/10.1016/j.rser.2013.03.004.

22. Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

23. Candanedo, Luis. (2016). Occupancy Detection [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5X01N.

24. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

25. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

26. Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

27. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

28. Matthews, B.W.. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. https://doi.org/10.1016/0005-2795(75)90109-9.

29. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.

30. Candanedo, Luis M.; Feldheim, Véronique. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO 2 measurements using statistical learning models. Energy and Buildings. https://doi.org/10.1016/j.enbuild.2015.11.071.