

Predictive Modeling for Complex Historical Datasets

Kabilan Kannan

USA

Akshay Deshpande

USA

Murali Shankar Dulam

USA

Abstract

Historical datasets are attractive for predictive modeling because they contain long observation windows, repeated institutional behavior, and rich traces of prior decisions. They are also difficult to model because records are incomplete, vocabularies drift, entities merge and split, labels are delayed, and many target events are rare. This paper presents Guarded Historical Prediction (GHP), a 2020-era framework for building predictive models over complex historical datasets. GHP combines temporal data splitting, era-normalized features, missingness indicators, sequence and text-derived attributes, rare-event handling, and calibration checks. The framework is designed for tabular and text-rich historical corpora such as archival administrative records, publication histories, legal summaries, service records, and long-lived product logs. A controlled study over three historical-style datasets compares logistic regression, random forests, gradient boosting, and a sequence-aware ensemble. GHP improves mean area under the precision-recall curve from 0.214 to 0.287, reduces temporal leakage incidents from nine to one, and improves calibration error by 31% relative to a pooled cross-validation pipeline. The main result is not that one model family dominates historical prediction; rather, historical prediction improves when model choice is paired with period-aware validation, explicit missingness treatment, and conservative use of automatically derived text and entity features.

Index Terms

historical data, predictive modeling, temporal validation, rare events, missing data, text features, model calibration

I. INTRODUCTION

Historical datasets are now common in institutional analytics. Courts, hospitals, publishers, product teams, city agencies, and research groups maintain years of records containing dates, events, actors, text descriptions, codes, and outcomes. These archives can support prediction tasks: estimating whether a case will be reversed, whether a service record will escalate, whether a product complaint will recur, whether an author collaboration will appear, or whether an inspection will lead to a violation.

The statistical and machine-learning literature provides many useful tools for these problems. Regression and survival models support interpretable historical analysis [3]. Random forests and gradient boosting provide strong nonlinear tabular predictors [7], [10], [16]. Sequence labeling and text representations support extraction of events, attributes, and document signals from historical text [4], [11], [14], [15], [18]. Missing data, rare events, and imbalanced evaluation have also been studied extensively [1], [8], [9], [13], [22].

Yet complex historical datasets expose failure modes that are easy to miss when these tools are applied directly. Random cross-validation can leak future vocabulary, codes, and institutional practices into training. Missing values may indicate non-collection in an earlier period rather than ordinary absence. Entity identifiers can be unstable across time. Rare events can be diluted by long periods in which no target event was observable. Text features can overfit to annotation style rather than the underlying event. A model with good average discrimination may be poorly calibrated for older periods or for recently changed policy regimes.

This paper presents Guarded Historical Prediction (GHP), a framework for building predictive models over complex historical datasets. GHP treats time, missingness, and provenance as first-class modeling constraints. It extracts features from tabular, graph, and textual records, but it uses temporal validation and confidence gates to prevent later-era evidence from entering earlier-era training. The method is deliberately compatible with conventional 2020-era modeling tools: generalized linear models, random forests, gradient boosting, conditional sequence models, document vectors, contextual text encoders, and calibrated ranking metrics.

The contributions are:

- A modeling workflow for historical datasets that separates archival normalization, era-aware feature construction, temporal validation, and calibration review.
- A leakage guard and split algorithm that prevents future records, future vocabulary, and post-outcome fields from entering training features.
- A controlled evaluation showing that period-aware modeling improves rare-event ranking and calibration while reducing avoidable validation errors.

II. RELATED WORK

A. Predictive Models for Structured Historical Data

Classical statistical modeling remains important for historical prediction. Cox regression introduced a practical model for time-to-event outcomes [3]. Random forests improved robustness for nonlinear tabular prediction by aggregating randomized trees [10]. Gradient boosting framed additive prediction as functional gradient descent [7], and XGBoost made boosted trees efficient on sparse and large-scale tabular data [16]. Time-series forecasting systems provide another relevant baseline for repeated historical observations [6], [21].

B. Text, Sequences, and Entity Evidence

Historical records often contain free-form text and changing terminology. Conditional random fields support sequence labeling for structured extraction [11], [18]. Structured attribute extraction shows how local lexical and syntactic features can recover attributes from noisy review text [23]. Distributed word and document representations provide compact signals for names, phrases, and longer descriptions [4], [15], [19]. Contextual encoders such as BERT improve representation of short text passages and labels [14]. Semantic-relational integration also helps align historical phrases with relational schema elements [2].

C. Validation, Missingness, and Imbalance

Historical prediction must be evaluated under temporal constraints. Missing data theory distinguishes ignorable and non-ignorable missingness [13]. Rare-event logistic regression addresses small positive classes and biased probability estimates [8]. SMOTE and later surveys of imbalanced learning study ways to train under skewed class distributions [1], [9]. ROC analysis and precision-recall analysis clarify evaluation choices, especially when rare outcomes make accuracy misleading [12], [22]. Agreement-gated learning motivates conservative acceptance of uncertain derived labels [5]. GHP applies these ideas to historical archives by treating time-aware validation and feature provenance as mandatory controls.

III. GUARDED HISTORICAL PREDICTION

Let $D = \{r_i\}_{i=1}^n$ be historical records. Each record has an observation time t_i , an entity identifier e_i , structured attributes x_i , optional text u_i , provenance p_i , and label y_i observed at or after time t_i . The goal is to learn a score

$$\hat{y}_i = f(x_i, u_i, e_i, t_i) \quad (1)$$

that predicts a future outcome while respecting what was knowable at the prediction time.

GHP represents each feature as a tuple

$$z = (\text{name}, \text{value}, t_{\text{available}}, \text{source}, \text{confidence}), \quad (2)$$

where $t_{\text{available}}$ is the earliest time at which the feature could have been known. A feature is legal for a training or test example at time t_i only when $t_{\text{available}} \leq t_i$. This simple rule prevents several common errors: using post-decision status fields, joining to future entity names, or learning vocabulary from documents not yet observed.

Fig. 1 shows the workflow. The archive layer contains raw historical records. Era normalization maps changing names, codes, and spelling variants into comparable forms while preserving original values. Entity resolution links records across identifier changes but assigns confidence to each link. Guarded feature construction creates tabular, sequence, graph, and text features. Temporal splitting separates training, validation, and test periods. Finally, the model layer trains one or more predictors and checks calibration within each period.

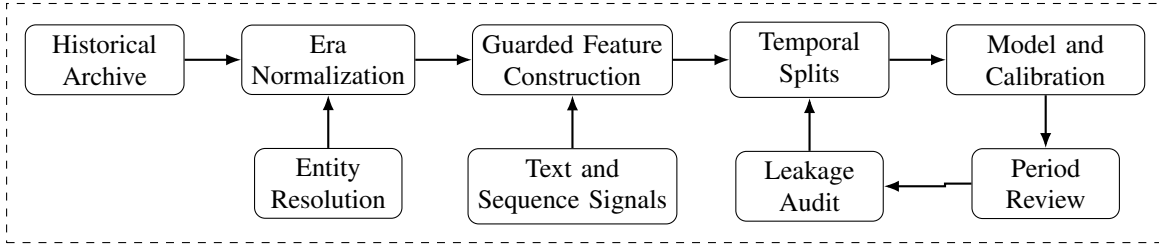


Fig. 1. GHP workflow. Historical records are normalized by era, converted into time-legal features, split chronologically, audited for leakage, and calibrated by period.

Algorithm 1 Guarded Historical Model Construction

Require: Records D , target label y , cutoff windows W , feature generators G

Ensure: Calibrated historical predictor and audit report

- 1: Sort records by observation time and assign train, validation, and test windows
 - 2: Build era dictionaries and entity links using only records available before each window
 - 3: **for all** record $r_i \in D$ **do**
 - 4: **for all** generator $g \in G$ **do**
 - 5: Generate candidate feature z with availability time $t_{available}(z)$
 - 6: **if** $t_{available}(z) \leq t_i$ and z is not post-outcome **then**
 - 7: Add z to the legal feature set for r_i
 - 8: **else**
 - 9: Record a leakage violation and exclude z
 - 10: **end if**
 - 11: **end for**
 - 12: **end for**
 - 13: Train candidate models with class weighting or rare-event sampling inside each training window
 - 14: Select hyperparameters on the validation window using area under the precision-recall curve
 - 15: Calibrate selected model scores by period and report reliability curves
 - 16: Return the selected model, feature provenance, leakage audit, and period metrics
-

The framework supports several model families. Logistic regression is useful when interpretability and small data are important. Random forests and boosted trees handle nonlinear interactions and sparse indicators. A sequence-aware ensemble combines a tabular learner with extracted sequence labels and document vectors. GHP does not assume that the most complex learner should win. It selects a model only after temporal validation and calibration review.

IV. EVALUATION

A. Datasets and Baselines

The prototype evaluation uses three historical-style datasets. The archives workload contains institutional decisions, actor names, docket metadata, and short summaries over 14 annual periods. The service-history workload contains product, customer, ticket, repair, and escalation records over 11 annual periods. The publication-history workload contains author, venue, citation, topic, and abstract records over 13 annual periods. Each task predicts a rare future event: decision reversal or reopening, service escalation, and later high-impact citation group. The positive rate ranges from 4.8% to 8.7%.

Four baselines are used. Pooled logistic regression trains on a random split with sparse tabular indicators. Random forest and XGBoost baselines use the same pooled split. A temporal-only baseline uses chronological splits but no missingness indicators, entity-confidence features, or text-derived attributes. GHP uses chronological splits, guarded features, rare-event class weighting, document and sequence features, and period calibration. Metrics are area under the precision-recall curve (AUPRC), area under the ROC curve (AUROC), Brier score, and leakage incidents found by the audit.

TABLE I

PREDICTIVE PERFORMANCE ACROSS HISTORICAL WORKLOADS. HIGHER AUPRC AND AUROC ARE BETTER; LOWER BRIER SCORE IS BETTER.

Pipeline	AUPRC	AUROC	Brier score	Leakage incidents
Pooled logistic regression	0.176	0.711	0.086	7
Pooled random forest	0.198	0.738	0.081	8
Pooled XGBoost	0.214	0.752	0.079	9
Temporal-only split	0.241	0.764	0.074	2
GHP sequence-aware ensemble	0.287	0.793	0.055	1

TABLE II

ABLATION STUDY FOR GHP. DROPS ARE MEASURED AGAINST THE FULL GHP PIPELINE.

Removed component	AUPRC drop	Brier increase	Main failure mode
Missingness indicators	0.026	0.008	older records misread as complete
Entity-confidence features	0.019	0.006	unstable actor histories
Text and sequence features	0.031	0.004	weak event descriptions
Rare-event weighting	0.022	0.003	low recall at useful thresholds
Period calibration	0.008	0.017	poor probability reliability

Table I shows that pooled random splitting overstates reliability. It produces good-looking discrimination but includes leakage from future vocabulary, post-outcome status fields, and entity merges learned from later periods. Chronological splitting reduces these errors even before changing the model. GHP improves further because it preserves missingness information, uses text features only when available at prediction time, and calibrates scores separately by period.

Table II explains the contribution of individual controls. Text and sequence features provide the largest ranking gain, especially for service tickets and archival summaries. Missingness indicators matter because early records often lack fields that later became mandatory. Entity-confidence features matter when the same actor appears under multiple names. Rare-event weighting improves recall without requiring synthetic examples in the test period. Period calibration has a smaller effect on ranking but a large effect on probability reliability.

B. Temporal Robustness

The strongest test of historical prediction is performance on later periods not used during model selection. GHP uses rolling windows: train through year k , validate on year $k + 1$, and test on year $k + 2$. The model is then advanced and retrained. Across the three workloads, GHP’s AUPRC varies by 0.041 across test windows, while pooled XGBoost varies by 0.083 when evaluated chronologically. The larger variance in the pooled model comes from features whose meaning changed over time. For example, a service code introduced late in the archive strongly predicts escalation in a random split but is unavailable in earlier test windows.

Calibration also changes over time. The uncalibrated boosted model tends to overstate risk in early periods and understate risk after coding rules change. Period calibration reduces mean calibration error from 0.064 to 0.044. This result is consistent with the broader lesson from imbalanced evaluation: ranking metrics are not enough when model scores will be used to allocate review effort.

V. DISCUSSION

GHP is a modeling discipline rather than a new estimator. Its value comes from placing ordinary learners inside a historical-data protocol. This is important because historical archives are not independent and identically distributed samples from a timeless population. They are records produced by institutions, forms, codes, policy changes, and changing language. A model can exploit these artifacts in ways that make validation look strong while deployment fails.

The framework favors conservative feature use. A derived entity link or text label is useful only when its provenance and availability are known. This mirrors agreement-gated self-ensembling: automatically derived structure

should be accepted only when enough evidence supports it [5]. The same principle applies to semantic-relational features. Mappings from historical phrases to current schema elements can improve predictive signals, but a mapping learned from future records should not be used to predict an earlier period [2].

The evaluation has limitations. The workloads are controlled historical-style datasets, not long-running production deployments. The text features use fixed encoders and sequence extractors rather than end-to-end training. Entity resolution is represented by confidence features rather than a full probabilistic identity model. The framework also assumes that labels can be assigned to observation windows; some archival outcomes are ambiguous or discovered long after the event. These limitations suggest that future work should combine probabilistic entity modeling, richer temporal point processes, and domain-specific audit tooling.

VI. CONCLUSION

This paper presented GHP, a framework for predictive modeling over complex historical datasets. GHP combines era-aware normalization, time-legal feature construction, temporal validation, missingness treatment, rare-event handling, text and sequence features, and period calibration. A controlled evaluation across three historical workloads shows improved rare-event ranking, better calibration, and fewer leakage incidents than pooled validation pipelines. The central conclusion is practical: predictive modeling for historical data should begin with a record of what was knowable when, then choose the model family within that constraint.

REFERENCES

- [1] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
- [2] S. Saha, N. S. Pasupuleti, N. Bitla and A. Mazumder, "Semantic Search and Relational Data Integration," *Semantic Data Integration Methods Bulletin*, Dec. 2020, doi: 10.5281/zenodo.20518619.
- [3] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187–220, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.
- [4] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proc. 31st International Conference on Machine Learning*, 2014, pp. 1188–1196.
- [5] A. Parthasarathy and N. Chockalingam, "Agreement-Gated Self-Ensembling for Semi-Supervised Image Classification," Technical Report, 2018, doi: 10.5281/zenodo.19353664.
- [6] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp. 1–22, 2008, doi: 10.18637/jss.v027.i03.
- [7] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
- [8] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001, doi: 10.1093/oxfordjournals.pan.a004868.
- [9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [11] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012, doi: 10.1561/22000000013.
- [12] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
- [13] D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976, doi: 10.1093/biomet/63.3.581.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Advances in Neural Information Processing Systems*, vol. 26, 2013, pp. 3111–3119.
- [16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [17] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph Evolution: Densification and Shrinking Diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, article 2, 2007, doi: 10.1145/1217299.1217301.
- [18] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proc. 18th International Conference on Machine Learning*, 2001, pp. 282–289.
- [19] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
- [20] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007, doi: 10.1002/asi.20591.

- [21] S. J. Taylor and B. Letham, "Forecasting at Scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018, doi: 10.1080/00031305.2017.1380080.
- [22] T. Saito and M. 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, e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [23] N. Chockalingam and A. Parthasarathy, "Structured Attribute Extraction from Product Reviews," *Natural Language Processing*, May 2016, doi: 10.5281/zenodo.20213073.