



A Review of Tabular Data Understanding and Automated Analytics

Priyanka Parate¹, Kapur Pardhi¹, Sanskruti Raut¹, Ayan Sheikh¹, Mrs. Umme Ayeman Saqib Gani²

¹Student, Department of Artificial Intelligence & Data Science Engineering, Priyadarshini College of Engineering, Nagpur (R.T.M. Nagpur University), Maharashtra, India

²Guide, Assistant Professor, Department of Artificial Intelligence & Data Science Engineering, Priyadarshini College of Engineering, Nagpur (R.T.M. Nagpur University), Maharashtra, India

ABSTRACT :

Reliable analysis of tabular data requires decisions about column meaning, data quality, model suitability and the interpretation of results. Progress in predictive modeling alone does not resolve these dependencies. This narrative review synthesizes 36 selected works spanning data profiling, semantic type detection, automated machine learning, tree ensembles, neural predictors, tabular foundation models, natural language interfaces, visualization and explainability. The comparison shows that tree ensembles remain useful baselines, while improved neural defaults, efficient ensembling and in-context learning broaden the available accuracy–cost trade-offs. Recent work on feature engineering and uncertainty also shows why predictive rankings cannot stand in for complete workflow evaluation. Existing AutoML systems already automate substantial parts of preparation and model construction; the unresolved challenge is to connect these capabilities to explicit semantic evidence, validated user intent and auditable explanations. The review identifies evaluation requirements for such integration without claiming a universally superior model or experimentally validating a new system.

Keywords: tabular data; semantic profiling; automated machine learning; foundation models; conversational analytics; explainable artificial intelligence.

1 Introduction

Tabular analytics supports decisions in business and scientific settings, but a table is not self-explanatory. An integer column may represent a measurement, a category or an identifier; repeated records may indicate duplication or legitimate events. The business intelligence framework of Chen et al. [1] places analysis within an organizational decision process. This perspective remains useful because a technically accurate prediction can still answer the wrong question when the dataset, target or aggregation has been misunderstood.

Profiling research describes how to extract structural metadata [2], while automated machine learning (AutoML) addresses the selection and configuration of learning pipelines [3]. These traditions have increasingly intersected with semantic typing and pretrained models. Nevertheless, they evaluate different outputs: a semantic label, a validation score, a generated query or an explanation. Treating all of these as interchangeable evidence of “intelligence” obscures their distinct failure modes.

This review examines how the selected literature supports a workflow from data understanding to prediction and interactive interpretation. Its contribution is a cross-stage synthesis: it compares what each family automates, the assumptions that remain external to it, and the evidence needed to connect the stages responsibly. The discussion concerns the reviewed literature rather than the measured performance of a new implementation.

2 Review Scope and Synthesis Method

The review uses a purposively selected corpus of 36 distinct sources, including an edited book and a book chapter. The seed bibliography combines predictive-learning studies with work on profiling, semantic interpretation and interactive analytics; overlapping entries were deduplicated. The publication span is 2000–2026. Bibliographic checking and targeted source retrieval were undertaken on 15 October 2026 using publisher proceedings, author or institutional pages, ACL Anthology, PMLR, Nature and arXiv. Exact titles and author names were used to resolve inconsistencies. This is a purposive narrative review, not an exhaustive systematic search or meta-analysis.

Sources were retained when they addressed profiling, semantic interpretation, preparation, predictive learning, model selection, interaction, visualization or reliability. Foundational NLP studies were included as enabling technologies, with their original evaluation domains kept explicit. Published versions were preferred where identifiable; arXiv identifiers identify preprint versions used for recent work. Available original text and official abstracts informed the synthesis; where full text was inaccessible, discussion was limited to the stated contribution and scope. No pooled effect sizes, new benchmark results or systematic-review screening counts are asserted.

Each source was examined along four axes: the task it automates, its mechanism, the scope of its reported evidence, and the decisions it leaves to the analyst. Differences in datasets, preprocessing, tuning budgets, hardware and metrics preclude a common numerical leaderboard. The analysis therefore

compares capabilities and qualifications rather than combining incompatible performance claims.

3 Data Understanding and Preparation

3.1 Structural profiling and semantic interpretation

Abedjan et al. [2] distinguish column statistics from more demanding discoveries such as unique combinations and dependencies. These observations establish what values and relationships occur; they do not establish whether a transformation is appropriate. For example, high uniqueness can suggest an identifier without proving that the column lacks predictive relevance. A preparation policy should therefore separate observed properties from inferred meaning and from actions taken on the data.

Sherlock [4] learns semantic column types from statistical, character and embedding features. Sato [5] adds table context through topic modeling and structured prediction. Their progression illustrates why isolated column inspection can be inadequate: neighboring fields provide evidence about an otherwise ambiguous column. Both methods are constrained by their training corpora and supported type vocabularies. A semantic label is consequently useful evidence for cleaning or schema matching, but should not be treated as an infallible domain judgment.

3.2 Feature engineering and task dependence

TabPrep [6] studies targeted feature generation and reports gains across multiple model families on TabArena. Its result challenges evaluations that attribute improvement solely to architecture while holding an impoverished feature representation fixed. Feature engineering, however, differs from correcting invalid records. Supervised encodings and other learned transformations must be fitted within training folds. The synthesis implication is that preparation should be assessed jointly with the downstream learner, while preserving a record of what changed and why.

4 Predictive Models and Transfer Across Tables

4.1 Tree ensembles as practical baselines

Random Forests [7] combine randomized trees, with internal error estimates and variable importance providing practical diagnostic tools. XGBoost [8] emphasizes scalable boosted learning, including sparsity-aware computation and weighted quantile sketches. LightGBM [9] introduces gradient-based sampling and exclusive feature bundling to reduce training cost. CatBoost [10] addresses categorical processing and prediction shift through ordered procedures. These contributions solve different statistical and computational problems; their original results do not establish a permanent ordering across all datasets.

The benchmark of Grinsztajn et al. [11] found strong tree performance on typical medium-sized tables and linked neural weaknesses to irrelevant features, feature orientation and irregular target functions. That finding is conditional on the benchmark and contemporary models. It supports retaining competitive tree baselines, rather than concluding that later neural or pretrained methods cannot improve on them. CatBoost illustrates how algorithm design can reduce the need for manual categorical encoding.

4.2 Neural models and stronger defaults

FT-Transformer [12] uses feature tokenization and attention, while the same study establishes a strong residual-network baseline under comparable protocols. RealMLP [13] shows the importance of architecture and meta-tuned defaults, with evaluation on datasets separated from those used to choose defaults. TabM [14] develops parameter-efficient ensembling with shared weights. Together, these studies weaken the assumption that tabular deep learning must rely on elaborate attention architectures. Their practical value depends on preprocessing, training budget and the diversity gained per unit of computation.

4.3 Pretraining and in-context prediction

XTab [15] transfers shared transformer components across tables with inconsistent schemas, using independent featurizers and federated pretraining. It remains distinct from in-context prediction, in which labeled rows are supplied as context without updating model parameters for each task. TabPFN [16] demonstrates the latter approach on small datasets using synthetic pretraining. Its reported scope of up to 10,000 samples and 500 features is a property of the studied version, not a limit that should be assigned to every subsequent tabular foundation model.

TabICL [17] develops an architecture for larger classification contexts. TabICLv2 [18] extends the approach to classification and regression with revised synthetic priors, attention and optimization. ConTextTab [19] integrates semantic representations and real-world pretraining into table-native in-context learning. These approaches address different constraints: context scale, prior diversity and access to column meaning. Their benchmark successes still require examination of pretraining overlap, hardware, schema compatibility and the behavior of rare or unfamiliar domains.

CRUMB [20] tackles inference cost by selecting distributionally matched contexts for batches of queries. As an inference wrapper, it should be distinguished from a newly pretrained predictor. Do-PFN [21] instead targets causal effects using priors over simulated causal structures. Its causal scope differs from ordinary supervised prediction, and its conclusions depend on the assumptions represented by those priors.

5 Automated Selection and Evaluation

5.1 Selection from experience and optimization

Landmarking [22] characterizes tasks using the performance of inexpensive learners, complementing descriptive dataset statistics. Vanschoren [23] places such reuse within the broader field of meta-learning. Both approaches offer a rationale for allocating computation to promising candidates, but transfer is useful only when past tasks represent the new setting. A recommendation based on dataset similarity should therefore initiate local validation rather than replace it.

Auto-WEKA [24] formulates combined algorithm selection and hyperparameter optimization. Auto-sklearn [25] brings together preprocessing search, experience from related datasets and ensembles of evaluated models. AutoGluon-Tabular [26] emphasizes multilayer stacking and bagging. These systems already integrate substantial preparation and modeling capabilities. Integration of preparation and model construction is therefore an established capability within the reviewed literature. A more precise question is whether automated choices are semantically justified, transparent to users and reliable under deployment constraints.

A domain study by Yu et al. [27] compares credit-risk pipelines combining models with dimensionality reduction and SMOTEENN. It illustrates the dependence of an observed winner on preparation and application context. Its ranking cannot be transferred to unrelated domains. More generally, resampling must occur inside the training portion of each split; otherwise synthetic examples can carry information across evaluation boundaries. Domain-specific evidence should be interpreted alongside broader benchmarks rather than substituted for them.

5.2 What benchmarking can establish

TabArena [28] provides a maintained evaluation framework and highlights the influence of validation, tuning budgets and ensembling. Its findings help reconcile apparently inconsistent claims about trees and neural models: different resource allocations can favor different families. Because a living benchmark changes, reproducibility requires recording the data, code, model and protocol versions used. A public ranking is evidence about its evaluated configuration, not an automatic decision rule for a new table.

Table 1 Comparison of predictive and selection approaches

Family and sources	Primary advantage	Qualification for use
Tree ensembles [7–11]	Strong task-specific learning and efficient baselines	Encoding, missing-value support and tuning differ by method
Neural predictors [12–14]	Learned interactions and efficient ensemble designs	Training and preprocessing budgets affect comparisons
Pretrained models [15–20]	Transfer across tables or rapid in-context prediction	Prior coverage, context cost and semantics remain version dependent
AutoML and meta-learning [3, 22–26]	Allocate search and combine pipeline components	Require an objective, valid splits and representative experience

For selection, the useful outcome is a justified accuracy–cost trade-off. Evaluation should include predictive loss, training and inference cost, uncertainty, and performance on relevant subgroups. The preferred predictor may change when latency or reliability matters more than a small average score gain.

6 Natural Language Interaction and Explainability

6.1 From language capability to validated analysis

The Transformer [29] establishes attention-based sequence modeling, and Brown et al. [30] demonstrate broad few-shot language capabilities. These are enabling results from language research, not direct evaluations of autonomous tabular analysis. A fluent response does not prove that the system selected the correct measure, respected units or executed a valid aggregation. Connecting language models to tables therefore introduces a separate grounding and verification problem.

Spider [31] makes schema generalization central to text-to-SQL evaluation by separating databases across training and testing. Its contribution is particularly relevant to unfamiliar tables: recognizing query wording is insufficient when column meanings and relationships change. Yet SQL execution alone does not resolve business ambiguity. A request for “best performance” may require clarification of the measure, direction and comparison period even when several executable queries are available.

Retrieval-augmented generation (RAG) [32] combines a generator with retrieved evidence. Its original evaluation concerns knowledge-intensive NLP, so improved tabular reasoning cannot be assumed from that result. In an analytical workflow, retrieved dictionaries and definitions can help interpret terminology, while numerical answers should be grounded in executed computations. Retrieval quality, source freshness and whether the cited evidence actually supports the answer require separate assessment. A retrieved passage is not a substitute for the current dataset.

6.2 Visualization and the meaning of explanation

Voyager [33] recommends charts through mixed-initiative exploration using statistical and perceptual considerations. This provides a useful precedent for systems that assist rather than silently decide. A generated visualization still requires valid variable roles and aggregation choices. Chart recommendation, question answering and predictive modeling solve related but different tasks, and success at one should not be reported as evidence for the others.

LIME [34] fits an interpretable surrogate around a prediction. SHAP [35] provides an additive feature-attribution framework with theoretically motivated properties. Both explain aspects of model behavior, with interpretation depending on the explanatory setup. Correlated features, reference distributions and local sampling choices can change the account given to the analyst. Attribution does not establish causation, and a readable explanation does not certify that a model is accurate or that its use is appropriate.

6.3 Uncertainty as a separate evaluation axis

De Melo Costa et al. [36] compare predictive discrimination with conformal uncertainty measures and report weaker size-stratified coverage for the evaluated foundation models despite high AUC. This result is bounded by the studied models, classification tasks and protocol. It should not be generalized to every later model. It nevertheless demonstrates why accuracy, uncertainty and explanation must be assessed separately. Marginal conformal coverage under exchangeability also does not guarantee coverage for each subgroup or under distribution shift.

A defensible interface should distinguish a computed statistic, a model attribution and a generated interpretation. This separation makes discrepancies visible and permits the system to ask for clarification or withhold an unsupported answer.

7 Research Gaps and Future Directions

The reviewed evidence suggests that integration should be evaluated as a chain of decisions. Semantic typing can influence feature exclusion; feature exclusion can change the fitted model; the model can change the explanation presented to a user. Measuring each stage independently misses these propagated effects. The directions in Table 2 are synthesis proposals arising from the corpus, rather than claims that no related systems exist.

Table 2 Research priorities and corresponding evaluation requirements

Priority	Unresolved question	Evidence needed
Semantic preparation	Do inferred roles justify transformations without removing useful meaning?	Role accuracy, transformation provenance and downstream ablations
Adaptive model choice	When does semantic evidence improve recommendations beyond statistical metadata?	Held-out task families, selection regret and compute-matched baselines
Grounded interaction	Does a correct query also capture the intended analytical question?	Unseen schemas, execution correctness and ambiguity-resolution tests
Reliable interpretation	Do explanations and confidence remain useful on difficult populations?	Attribution stability, calibration and subgroup or shift evaluation
Workflow integration	Does combining components improve user decisions?	Stage-level failure analysis and controlled user studies

Semantic understanding should be treated as revisable evidence. Systems could retain multiple candidate roles when confidence is low, combine column values with documentation and table context, and request confirmation when a choice affects the target or an irreversible transformation. Experiments should test whether this additional context improves outcomes over simpler rules, while reporting the cost of clarification and the rate of unnecessary interventions.

Model recommendation should account for data scale, categorical structure, available compute and the intended use. A useful experiment would compare metadata-only selection, semantic enrichment and a fixed strong baseline under the same resource budget. Foundation-model variants should be identified explicitly because changes in priors, training data or context handling can materially alter applicability. Preparation methods such as TabPrep [6] should be included in these comparisons rather than treated as unrelated additions.

Interactive workflows also need evaluation beyond polished demonstrations. A benchmark should include ambiguous requests, unsupported causal questions, stale metadata and tables with overlapping column meanings. It should record when the system correctly asks a question, abstains or supplies an answer with traceable computation. Human studies can then examine whether these behaviors improve decisions, rather than measuring satisfaction alone.

8 Conclusion

The selected literature supports a plural approach to tabular analytics. Tree ensembles, improved neural models, pretrained predictors and AutoML systems offer different strengths, and their usefulness depends on data characteristics and evaluation budgets. Profiling and semantic typing provide evidence about tables; language interfaces and visualization broaden access; attribution and uncertainty methods expose different aspects of reliability. The main opportunity is to connect these capabilities through explicit assumptions and verifiable computations. Progress should be demonstrated through

reproducible comparisons and analysis of errors across stages, with claims restricted to the settings actually evaluated.

REFERENCES

- [1] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.
- [2] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: A survey," *The VLDB Journal*, vol. 24, pp. 557–581, 2015.
- [3] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019.
- [4] M. Hulsebos et al., "Sherlock: A deep learning approach to semantic data type detection," in *Proc. KDD*, pp. 1500–1508, 2019.
- [5] D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W.-C. Tan, "Sato: Contextual semantic type detection in tables," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 1835–1848, 2020.
- [6] A. Tschalzev et al., "TabPrep: Closing the feature engineering gap in tabular benchmarks," *arXiv:2606.02384*, 2026, preprint version.
- [7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, pp. 785–794, 2016.
- [9] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [10] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [11] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [12] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [13] D. Holzmüller, L. Grinsztajn, and I. Steinwart, "Better by default: Strong pre-tuned MLPs and boosted trees on tabular data," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [14] Y. Gorishniy, A. Kotelnikov, and A. Babenko, "TabM: Advancing tabular deep learning with parameter-efficient ensembling," in *Proc. ICLR*, 2025.
- [15] B. Zhu, X. Shi, N. Erickson, M. Li, G. Karypis, and M. Shoaran, "XTab: Cross-table pretraining for tabular transformers," in *Proc. ICML, PMLR*, vol. 202, pp. 43181–43204, 2023.
- [16] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, 2025.
- [17] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICL: A tabular foundation model for in-context learning on large data," in *Proc. ICML, PMLR*, vol. 267, pp. 50817–50847, 2025.
- [18] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICLv2: A better, faster, scalable, and open tabular foundation model," in *Proc. ICML, PMLR*, vol. 306, pp. 102477–102555, 2026.
- [19] M. Spinaci, M. Polewczyk, M. Schambach, and S. Thelin, "ConTextTab: A semantics-aware tabular in-context learner," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [20] J. Heredge, M. J. Villani, P. Deshpande, A. Seshadri, and N. Kumar, "CRUMB: Efficient prior fitted network inference via distributionally matched context batching," *arXiv:2606.11473*, 2026, preprint.
- [21] J. Robertson, A. Reuter, S. Guo, N. Hollmann, F. Hutter, and B. Schölkopf, "Do-PFN: In-context learning for causal effect estimation," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [22] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier, "Meta-learning by landmarking various learning algorithms," in *Proc. ICML*, pp. 743–750, 2000.
- [23] J. Vanschoren, "Meta-learning," in *Automated Machine Learning*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., pp. 35–61, 2019.
- [24] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proc. KDD*, pp. 847–855, 2013.
- [25] M. Feurer et al., "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, vol. 28, pp. 2962–2970, 2015.
- [26] N. Erickson et al., "AutoGluon-Tabular: Robust and accurate AutoML for structured data," *arXiv:2003.06505*, 2020.
- [27] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng, "Advanced user credit risk prediction model using LightGBM, XGBoost and Tabnet with SMOTEENN," *arXiv:2408.03497*, 2024.
- [28] N. Erickson et al., "TabArena: A living benchmark for machine learning on tabular data," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [29] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
- [30] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

-
- [31] T. Yu et al., “Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task,” in Proc. EMNLP, pp. 3911–3921, 2018.
- [32] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
- [33] K. Wongsuphasawat et al., “Voyager: Exploratory analysis via faceted browsing of visualization recommendations,” IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 649–658, 2016.
- [34] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proc. KDD, pp. 1135–1144, 2016.
- [35] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
- [36] J. L. De Melo Costa, F. Popineau, A. Rimmel, and B.-L. Doan, “High performance, low reliability: Uncertainty benchmarking for tabular foundation models,” in Proc. ESANN, pp. 115–120, 2026.