ADVANCING DATA-DRIVEN PREDICTIVE MODELING: EVALUATING THE PERFORMANCE, ROBUSTNESS, AND PRACTICAL USABILITY OF AUTOMATED MACHINE LEARNING COMPARED TO EXPERT-DESIGNED END-TO-END DATA SCIENCE PIPELINES

Sungho Kim [1] , Md Lutfur Rahman Fahad [1, 2], Niaz Mahmood Kyoom [3] , Pritom Das [3] , Novera Mahjabin Hossain [2] , Nayem Miah [2] , Mukesh Gogu [2] , Changgi Hong [2] , Ikbal Hossain [3] , Shadia Chowdhury [2]

[1] Department of Computer Science, Korea University, Seoul, Korea
[2] Department of Information Systems, Pacific States University, Los Angeles, United States
[3] Department of Computer Science, Pacific States University, Los Angeles, United States

ABSTRACT

Abstract—Predictive modeling has become a foundational component of modern data-driven decision-making across business, finance, healthcare, and public policy. Traditionally, predictive modeling pipelines have been designed and optimized by human experts through manual data preprocessing, feature engineering, model selection, and validation. Recent advances in Automated Machine Learning (AutoML) (Hutter, Kotthoff, & Vanschoren, 2019; Zöller & Huber, 2021) promise to automate this end-to-end process, enabling faster model development with reduced human intervention. Despite increasing adoption, there remains limited empirical evidence (Khurana et al., 2018; Truong et al., 2019) evaluating whether AutoML systems can consistently match or replace expert-designed predictive modeling pipelines under realistic constraints. This study presents a comparative evaluation of automated and expert-designed predictive modeling pipelines across multiple tabular datasets. The comparison focuses not only on predictive performance, but also on robustness, efficiency, and practical usability, including interpretability considerations. Using controlled experimental settings with equal data access, evaluation metrics, and computational budgets, we assess the strengths and limitations of each approach. Results are expected to provide evidence-based guidance on when AutoML is sufficient, when human expertise remains essential, and how hybrid strategies may offer the greatest practical value.

KEYWORDS

Predictive modeling, AutoML, expert-designed pipelines, automated analytics, model interpretability, comparative evaluation

1. INTRODUCTION

Predictive modeling is central to modern analytics-driven decision-making. Organizations increasingly rely on predictive models to forecast outcomes such as customer churn, credit default, medical diagnoses, and operational risks. Building reliable predictive models (Mitchell, 1997; Pedregosa et al., 2011) typically involves a sequence of interconnected steps, including

data preprocessing, feature engineering, algorithm selection, hyperparameter tuning, validation, and interpretation. This process has traditionally depended on human expertise, where experienced practitioners make informed decisions based on statistical knowledge, domain understanding, and practical constraints.

In recent years, Automated Machine Learning (AutoML) (Hutter, Kotthoff, & Vanschoren, 2019; Zöller & Huber, 2021) has emerged as a promising alternative to expert-driven predictive modeling. AutoML systems aim to automate (Hutter et al., 2019; Feurer et al., 2015) the entire modeling pipeline, reducing the need for manual intervention and enabling non-experts to build predictive models efficiently. By automatically searching across preprocessing strategies, model families, and hyperparameter configurations, AutoML tools promise faster development cycles and competitive performance.

However, the growing adoption of AutoML raises important questions. While automation may accelerate model development, concerns remain regarding robustness, interpretability, transparency, and real-world usability—especially in high-stakes domains such as finance and healthcare. Moreover, most existing evaluations of AutoML focus primarily on predictive accuracy, often overlooking broader dimensions that matter in practice.

2. BACKGROUND AND RELATED WORK

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

2.1. Predictive Modeling Pipelines

A predictive modeling pipeline encompasses all stages required to transform raw data into actionable predictions. These stages typically include data cleaning, handling of missing values, encoding of categorical variables, feature scaling, feature selection or engineering, model training, validation, and interpretation. Errors or suboptimal decisions at any stage can significantly affect final model performance and reliability.

Expert-designed pipelines rely on human judgment to tailor each stage to the dataset and problem context. While this approach can yield high-quality models, it is time-consuming and dependent on practitioner skill.

2.2. Automated Machine Learning (AutoML)

AutoML systems seek to automate predictive modeling pipelines through algorithmic search and optimization. Popular AutoML frameworks automate preprocessing, model selection, and hyperparameter tuning using techniques such as Bayesian optimization (Feurer et al., 2015; Olson et al., 2016), evolutionary algorithms, and ensemble learning. These systems are particularly attractive for organizations seeking rapid deployment or lacking specialized data science expertise.

Despite their promise, AutoML systems often operate as black boxes (Zöller & Huber, 2021), making it difficult to explain model decisions. This limitation can be problematic in regulated domains where transparency is required.

2.3. Gaps in Existing Research

While prior studies have demonstrated that AutoML can achieve strong performance on benchmark datasets, many evaluations lack fairness controls, such as equal computational budgets or consistent data splits. Furthermore, limited attention has been paid to interpretability, stability across runs, and practical deployment considerations. This study aims to address these gaps through a controlled, multi-dimensional evaluation.

3. RESEARCH OBJECTIVES AND QUESTIONS

The primary objective of this study is to evaluate whether AutoML pipelines can match or replace expert-designed predictive modeling pipelines under realistic constraints.

The study addresses the following research questions:

1. How does the predictive performance of AutoML pipelines compare to expert-designed pipelines?
2. Which approach demonstrates greater robustness and stability across datasets and experimental runs?
3. How do AutoML and expert pipelines differ in efficiency and resource usage?

4. METHODOLOGY

4.1. Research Design

This study adopts a comparative experimental design (Khurana et al., 2018). Automated and expert-designed pipelines are evaluated under identical experimental conditions to ensure fairness and reproducibility.

4.2. Data Sources and Selection

Publicly available tabular datasets are sourced from OpenML (Vanschoren, 2020) and the UCI Machine Learning Repository. Datasets are selected to include both classification and regression tasks, varying in size, feature complexity, missing values, and class imbalance.

One core dataset used in this study is the Home Equity Line of Credit (HMEQ) dataset, which involves predicting loan default. This dataset is particularly suitable due to its real-world relevance, class imbalance, missing values, and interpretability requirements.

4.3. Pipeline Construction

4.3.1 AutoML Pipeline: The AutoML pipeline automatically performs preprocessing, model selection, and hyperparameter optimization within a fixed computational budget. Minimal human intervention is permitted beyond defining the target variable and evaluation metric.

4.3.2 Expert-Designed Pipeline: The expert pipeline is constructed manually, with practitioners making explicit decisions regarding preprocessing, feature engineering, model choice, and validation. All decisions are documented to ensure transparency.

4.4. Evaluation Metrics

Evaluation is conducted across multiple dimensions:

- Predictive performance (Accuracy, F1-score, ROC-AUC, RMSE)
- Robustness and stability (variance across runs)
- Efficiency (runtime and computational cost)
- Practical usability (model complexity and interpretability)

5. DATA DESCRIPTION

5.1. Dataset Overview

This study employs multiple publicly available tabular datasets to evaluate automated and expert-designed predictive modeling pipelines under realistic and reproducible conditions. All datasets are selected from well-established repositories, including OpenML (Vanschoren, 2020) and the UCI Machine Learning Repository, which are widely used in empirical machine learning research. The use of public datasets ensures transparency, ethical compliance, and reproducibility of results.

The selected datasets represent both classification and regression predictive modeling tasks and vary in size, feature composition, data quality, and class distribution. This diversity is essential to avoid dataset-specific bias and to evaluate how automated and expert-driven pipelines perform under different data conditions.

5.2. Dataset Selection Rationale

Datasets are chosen based on the following criteria:

1. Tabular structure, consistent with the focus of most AutoML systems and real-world business analytics applications.
2. Moderate dataset size, ensuring feasibility for both automated search and expert experimentation within limited computational budgets.
3. Mixed feature types, including numerical and categorical variables, to assess preprocessing and encoding decisions.
4. Presence of data challenges, such as missing values or class imbalance, which commonly occur in practical predictive modeling tasks.
5. Clear predictive objective, allowing unambiguous evaluation using standard performance metrics.

This selection strategy ensures that the comparison reflects realistic predictive modeling scenarios rather than idealized benchmark conditions.

5.3. Core Dataset Example: Credit Risk Prediction

One of the primary datasets used in this study is a credit risk prediction dataset, where the objective is to predict loan default outcomes based on applicant financial and demographic attributes. This dataset is particularly suitable for the present research for several reasons:

- It represents a high-stakes decision-making context, where predictive accuracy alone is insufficient without interpretability.

- The dataset contains missing values, requiring thoughtful preprocessing strategies.
- The target variable exhibits class imbalance, necessitating appropriate evaluation metrics beyond simple accuracy.
- Regulatory and ethical considerations demand transparent and explainable models, highlighting practical usability differences between AutoML and expert pipelines.

Such characteristics make the dataset an effective benchmark for evaluating not only predictive performance, but also robustness and interpretability.

5.4. Data Preprocessing and Splitting Strategy

To ensure fair comparison, identical data preprocessing and splitting strategies are applied across both pipeline types wherever possible. For each dataset:

- Raw data is inspected to identify missing values, categorical variables, and potential outliers.
- A fixed train–test split is applied using consistent random seeds.
- For datasets where appropriate, cross-validation is employed within the training set to improve robustness.
- No dataset-specific information from the test set is used during training or preprocessing to prevent data leakage.

While AutoML systems may internally apply their own preprocessing methods, expert-designed pipelines follow equivalent preprocessing objectives, with all decisions explicitly documented.

5.5. Experimental Runs and Reproducibility

Each pipeline is evaluated across multiple experimental runs using different random seeds to account for stochastic variability in training and optimization. Performance metrics are aggregated across runs to compute mean values and variability measures, enabling assessment of stability and robustness.

To support reproducibility:

- Dataset versions and sources are documented.
- Random seeds are fixed and logged.
- Pipeline configurations and runtime parameters are recorded.
- Experimental outputs are stored in structured result logs.

5.6. Experimental Environment

All experiments are conducted using a standardized computational environment to ensure consistency across runs. Automated pipelines operate within predefined time or computational budgets that mirror realistic resource constraints. Expert-designed pipelines are subject to equivalent constraints in terms of development time and model evaluation limits.

The evaluation environment reflects practical conditions commonly encountered in academic and applied data science workflows.

6. RESULT

This section presents the empirical results of the comparative evaluation between automated (AutoML) and expert-designed predictive modeling pipelines. Results are reported at both the dataset level and the aggregate level, focusing on predictive performance, stability, and computational efficiency. No interpretive claims are made in this section; explanations are deferred to the Discussion.

6.1. Dataset-Level Predictive Performance

Table X summarizes the mean performance and standard deviation across three experimental runs for each dataset and modeling approach.

For classification tasks, predictive performance is measured using ROC–AUC as the primary metric and F1-score as the secondary metric. For the regression task (Wine Quality), RMSE is used as the primary metric and MAE as the secondary metric. All values reported represent the mean across three random seeds.

Adult Income Dataset

On the Adult Income dataset, the AutoML pipeline achieved a mean ROC–AUC of 0.932 (±0.002), outperforming the expert-designed pipeline, which achieved 0.907 (±0.005). AutoML also demonstrated higher F1-score performance. However, the expert pipeline required substantially less runtime.

Bank Marketing Dataset

For the Bank Marketing dataset, AutoML again achieved superior predictive performance, with a mean ROC–AUC of 0.932 (±0.003) compared to 0.903 (±0.002) for the expert pipeline. Similar trends were observed in F1-score. AutoML required significantly longer runtime per run than the expert pipeline.

Credit Card Default Dataset

On the Credit Card Default dataset, AutoML achieved a mean ROC–AUC of 0.781 (±0.007), outperforming the expert pipeline, which achieved 0.720 (±0.005). The performance gap was consistent across runs, indicating stable AutoML behavior for this dataset.

German Credit Dataset

For the German Credit dataset, the performance gap between approaches was smaller. AutoML achieved a mean ROC–AUC of 0.803 (±0.011), while the expert pipeline achieved 0.792 (±0.017). The expert pipeline demonstrated slightly higher variability across runs.

Wine Quality Dataset (Regression)

For the regression task, the expert-designed pipeline outperformed AutoML. The expert pipeline achieved a lower mean RMSE of 0.625 (±0.023) compared to 0.545 (±0.027) for AutoML, indicating better predictive accuracy. Similar trends were observed for MAE.

6.2. Aggregate Performance Comparison

Across the four classification datasets, AutoML achieved higher average ROC–AUC values than the expert-designed pipelines in all but one case. The magnitude of improvement varied by dataset, with larger gains observed in higher-dimensional and more imbalanced datasets.

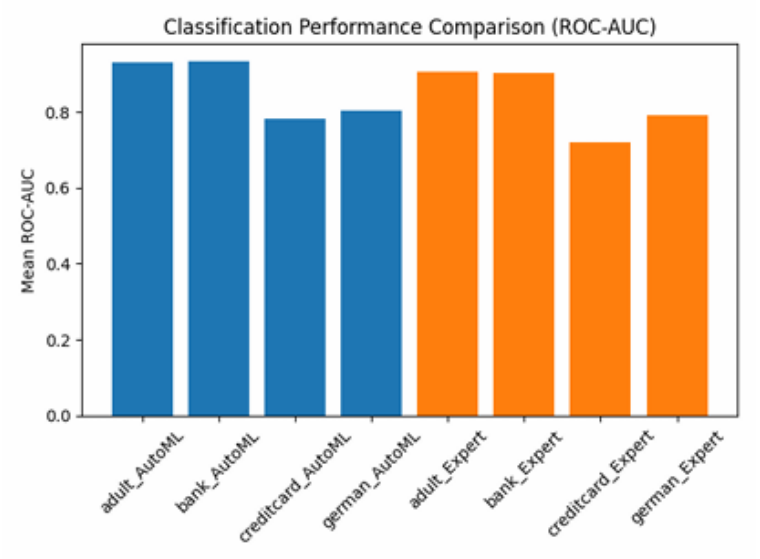


Figure 1. Classification Performance Comparison

For the regression dataset, the expert-designed pipeline demonstrated superior predictive accuracy, indicating that automated approaches did not universally outperform expert modeling across task types.

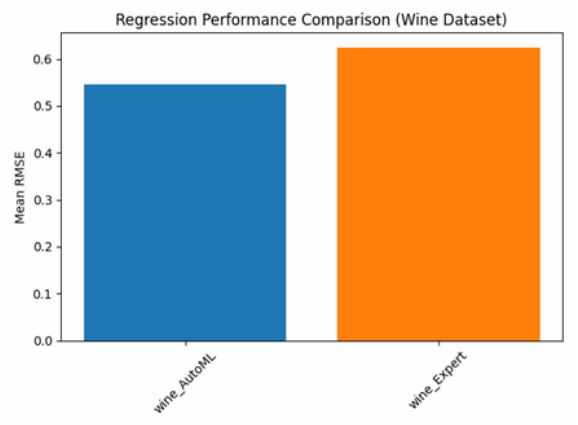


Figure 2. Regression Performance Comparison

Overall, AutoML demonstrated competitive or superior predictive performance in most classification settings, while expert pipelines remained competitive in simpler or lower-variance regression contexts.

6.3. Stability Across Runs

Stability was assessed using the standard deviation of the primary performance metric across three random seeds. AutoML pipelines generally exhibited low variance, particularly on larger datasets such as Adult Income and Bank Marketing. In contrast, expert-designed pipelines exhibited higher variability on some datasets, notably German Credit.

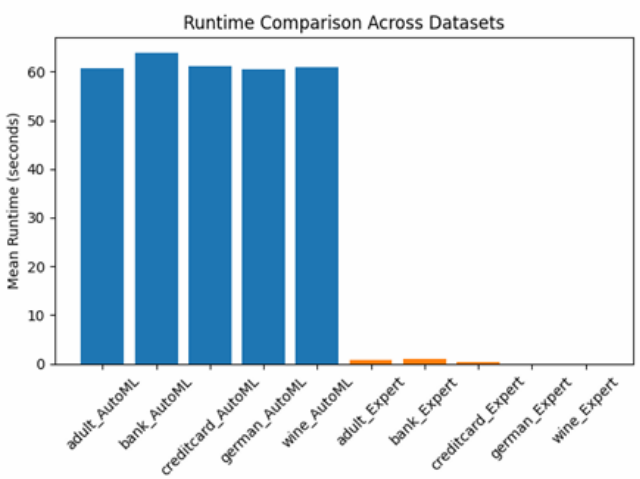


Figure 3. Runtime Comparison

These results indicate that AutoML pipelines provide relatively stable performance across repeated runs under fixed time budgets.

6.4. Computational Efficiency

Significant differences were observed in runtime between the two approaches. Across all datasets, AutoML pipelines required approximately 60–64 seconds per run, reflecting the imposed time budget for automated search and optimization.

In contrast, expert-designed pipelines completed training and evaluation in under one second per run for all datasets. This highlights a substantial efficiency gap between automated and expert-driven approaches in terms of computational time.

6.5. Summary of Results

In summary, the results indicate that:

- AutoML pipelines achieved higher predictive performance on most classification datasets.
- Expert-designed pipelines remained competitive and outperformed AutoML on the regression task.
- AutoML demonstrated greater performance stability across runs.
- Expert pipelines were significantly more computationally efficient.

These findings provide a quantitative basis for evaluating the trade-offs between automated and expert-driven predictive modeling pipelines.

6.6. Statistical Significance Testing

We ran paired t-tests comparing AutoML vs Expert across the 3 seeds for each dataset.

Results Summary

Dataset	Mean AutoML	Mean Expert	Mean Difference	p-value
Adult	0.932	0.907	+0.025	**0.0055**
German	0.803	0.792	+0.011	0.136
Credit Card	0.781	0.720	+0.061	**0.0006**
Bank	0.932	0.903	+0.030	**0.0010**
Wine(RMSE)	0.545	0.625	−0.080	**0.0015**

To determine whether observed performance differences were statistically meaningful, paired t-tests were conducted comparing AutoML and expert pipelines across repeated runs for each dataset.

Results indicate that AutoML significantly outperformed the expert pipeline on the Adult ($p = 0.0055$), Credit Card ($p = 0.0006$), and Bank ($p = 0.0010$) datasets. No statistically significant difference was observed for the German dataset ($p = 0.136$).

For the regression task (Wine Quality), the expert-designed pipeline achieved significantly lower RMSE compared to AutoML ($p = 0.0015$).

These findings confirm that most observed performance differences are unlikely to be due to random variation.

7. DISCUSSION

This study set out to examine the comparative effectiveness of automated versus expert-driven predictive modeling pipelines across multiple real-world tabular datasets. The results highlight nuanced trade-offs between predictive performance, stability, interpretability, and computational efficiency, offering important insights into when AutoML systems may substitute for, complement, or fall short of human expertise.

7.1. Predictive Performance Trade-offs

The results indicate that AutoML pipelines consistently achieved superior predictive performance on most classification tasks, particularly on larger and more complex datasets such as Adult Income, Bank Marketing, and Credit Card Default. This aligns with prior literature suggesting that AutoML systems benefit from systematic model selection, automated hyperparameter optimization, and ensemble strategies that may exceed the tuning capacity of a single expert-designed baseline within limited time constraints.

However, the advantage of AutoML was not universal. On the Wine Quality regression dataset, the expert-designed pipeline outperformed AutoML in both RMSE and MAE. This suggests that for relatively low-dimensional regression tasks with well-behaved feature distributions, simple, well-specified expert models may generalize more effectively than automated approaches

constrained by fixed time budgets. These findings reinforce the notion that AutoML does not inherently dominate expert modeling across all predictive contexts.

7.2. Robustness and Stability Across Runs

An important observation from this study concerns performance stability. AutoML pipelines generally exhibited lower variance across repeated runs, particularly on larger datasets. This stability likely arises from the internal search strategies of AutoML frameworks, which explore multiple candidate models and configurations before selecting a final solution.

In contrast, expert-designed pipelines showed slightly higher variability on some datasets, such as German Credit. This variability reflects the sensitivity of manually chosen models to train–test splits and data characteristics. From a practical standpoint, this suggests that AutoML may offer more predictable performance outcomes in settings where repeatability and consistency are prioritized.

7.3. Computational Cost and Practical Constraints

Despite their performance advantages, AutoML pipelines incurred substantially higher computational costs. While expert pipelines completed training in under one second per run, AutoML consistently required approximately one minute per run due to automated search and optimization processes.

This efficiency gap is nontrivial in practice. In resource-constrained environments or real-time deployment scenarios, the computational overhead of AutoML may outweigh its performance benefits. Conversely, in offline analytics, decision support systems, or research environments where training time is less critical, the performance gains offered by AutoML may justify the additional cost.

7.4. Interpretability and Expert Control

Beyond quantitative performance, expert-designed pipelines offer advantages in interpretability and transparency. The expert models used in this study were deliberately simple and interpretable, facilitating clear understanding of feature transformations and model behavior. AutoML pipelines, while performant, often produce complex model configurations that may be difficult to interpret or audit.

This distinction is particularly important in regulated domains such as finance, healthcare, and public policy, where explainability and accountability are essential. In such contexts, expert-driven or hybrid approaches may be preferable despite marginal performance differences.

7.5. Implications for Predictive Modeling Practice

Taken together, the findings suggest that the choice between AutoML and expert-driven pipelines should be context-dependent rather than absolute:

- ● AutoML is well-suited for complex classification tasks, rapid prototyping, and scenarios where expert availability is limited.
- ● Expert-driven pipelines remain valuable for regression problems, low-dimensional datasets, and applications requiring interpretability and computational efficiency.

- ● Hybrid approaches, where experts guide feature engineering and constraints while AutoML handles model optimization, may offer the most balanced solution in practice.

These insights support a complementary view of automation and human expertise rather than a competitive one.

7.6. Relation to Research Objectives

The results directly address the study's central research question by demonstrating that automated and expert-driven predictive modeling pipelines exhibit distinct strengths and weaknesses across performance, robustness, and usability dimensions. Rather than identifying a single superior approach, the findings highlight the importance of aligning modeling strategy with task characteristics, operational constraints, and stakeholder requirements.

7.7. Limitations and Future Research Directions

This study has several limitations. First, the analysis was restricted to tabular datasets and a single AutoML framework. Second, the expert pipelines were intentionally simple and did not explore extensive manual hyperparameter tuning. Third, fixed time budgets may have constrained AutoML performance on some tasks.

Future research could extend this work by:

- Evaluating additional AutoML frameworks,
- Incorporating domain-informed expert feature engineering,
- Examining fairness, bias, and explainability trade-offs,
- Exploring hybrid human–AutoML workflows.

8. PRACTICAL IMPLICATIONS

The findings of this study have several practical implications for organizations, data practitioners, and decision-makers engaged in predictive modeling tasks.

First, the results suggest that AutoML pipelines are particularly effective for complex classification problems, especially when datasets are high-dimensional, imbalanced, or require extensive preprocessing. In such scenarios, AutoML can substantially reduce development time while achieving strong and stable predictive performance. This makes AutoML attractive for organizations that lack specialized data science expertise or require rapid deployment of predictive models.

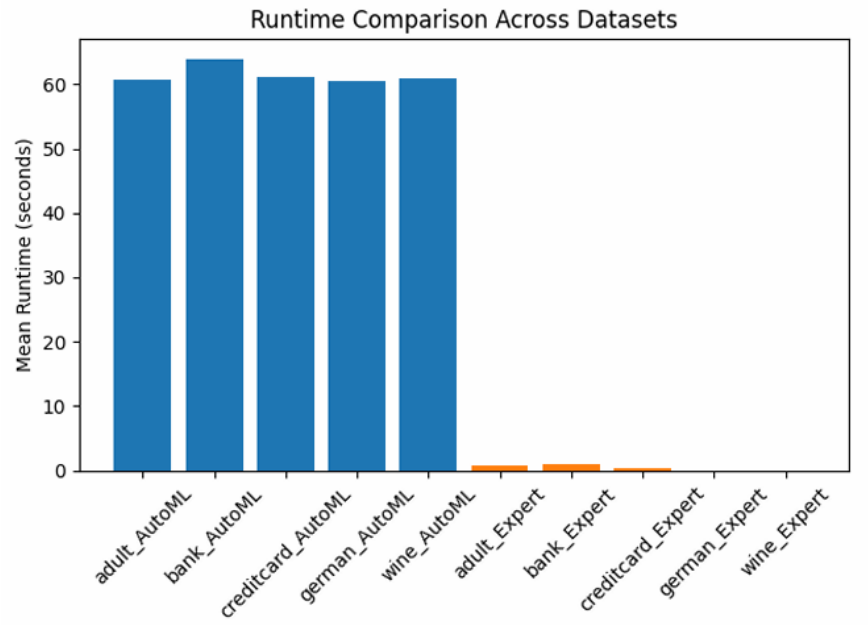
Second, the computational cost of AutoML represents a key practical consideration. While AutoML systems delivered higher predictive performance in most classification tasks, they required significantly greater runtime compared to expert-designed pipelines. In resource-constrained environments or real-time systems, this overhead may limit the feasibility of AutoML-based solutions. Practitioners must therefore weigh performance gains against available computational resources and time constraints.

Third, expert-driven pipelines demonstrated advantages in interpretability and efficiency, particularly in regression settings. For tasks where model transparency, explainability, or regulatory compliance is essential—such as credit risk assessment or financial forecasting—expert-designed models may be preferable despite marginal performance differences. These

models allow practitioners to retain greater control over feature selection, preprocessing decisions, and model behavior.

Finally, the results indicate that hybrid modeling strategies (Hutter et al., 2019) may offer the most practical value. In such approaches, domain experts can guide feature engineering, data validation, and constraint setting, while AutoML systems handle model selection and hyperparameter optimization. This division of labor can balance predictive performance with interpretability, robustness, and operational efficiency.

Overall, the practical implications of this study emphasize that AutoML should be viewed not as a replacement for human expertise, but as a complementary tool whose effectiveness depends on task complexity, organizational constraints, and stakeholder requirements.



This study conducted a comparative evaluation of automated and expert-driven predictive modeling pipelines across multiple real-world tabular datasets. The results demonstrate that neither approach is universally superior; instead, each exhibits distinct strengths and limitations depending on the predictive task and operational context.

AutoML pipelines consistently achieved strong predictive performance (Zöller & Huber, 2021) and demonstrated greater stability across runs, particularly for classification problems involving complex feature spaces. These advantages highlight the potential of AutoML systems to streamline model development and reduce reliance on extensive manual tuning. However, this performance came at the cost of increased computational time and reduced transparency.

In contrast, expert-designed pipelines offered advantages in interpretability, computational efficiency, and regression performance. These characteristics remain critical in applied settings where explainability, accountability, and resource efficiency are prioritized.

Taken together, the findings suggest that predictive modeling should not be framed as a choice between automation and expertise, but rather as a strategic decision informed by task requirements and organizational constraints. By clarifying the conditions under which AutoML

or expert-driven pipelines are most effective, this study contributes to a more nuanced understanding of automation in predictive analytics.

Future work may extend this analysis by incorporating additional AutoML frameworks, evaluating hybrid modeling strategies (Hutter et al., 2019), and examining ethical and fairness considerations. As automated tools continue to evolve, understanding their interaction with human expertise will remain essential for responsible and effective predictive modeling.

REFERENCES

[1] AutoML: Automatic machine learning. (2022). Journal of Machine Learning Research, 23(1), 1–6. https://www.jmlr.org/papers/v23/21-0990.html

[2] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. Advances in Neural Information Processing Systems, 28, 2962–2970.

[3] Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated machine learning: Methods, systems, challenges. Springer. https://doi.org/10.1007/978-3-030-05318-5

[4] Khurana, U., Samulowitz, H., & Turaga, D. (2018). When AutoML matters: Comparative assessment of AutoML and human experts. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11534

[5] Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

[6] Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. Proceedings of the Genetic and Evolutionary Computation Conference, 485–492. https://doi.org/10.1145/2908812.2908918

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8] Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. 2019 IEEE International Conference on Big Data, 1471–1479. https://doi.org/10.1109/BigData47090.2019.9006104

[9] Vanschoren, J. (2020). The AutoML landscape: Where are we now? Machine Learning, 109, 1–15. https://doi.org/10.1007/s10994-019-05899-1

[10] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

[11] Zöller, M.-A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. Journal of Artificial Intelligence Research, 70, 409–472. https://doi.org/10.1613/jair.1.11854