Comparação de Frameworks de AutoML: Desempenho e Seleção de Modelos

Bruno Chimentão Punhagui[1], **Alessandro Botelho Bovo**[2], **Danilo Sipoli Sanches**[1]

[1]Universidade Técnológica Federal do Paraná (UTFPR)
Cornélio Procópio, PR – Brazil

[2]Universidade Técnológica Federal do Paraná (UTFPR)
Londrina, PR – Brazil

brunochimentao,alessandrobovo,danilosanches@utfpr.edu.br

Abstract. *This study presents a comparative analysis of five AutoML frameworks applied to five binary classification datasets, evaluating model performance. The AUC metric was used as the main comparison criterion. Without ensembles, AutoGluon and MLJAR alternated in the top positions. With ensembles, techniques that combine multiple models to reduce errors and improve accuracy, AutoGluon ranked first in all datasets. In addition to superior performance.*

Resumo. *Este trabalho apresenta uma análise comparativa de cinco frameworks de AutoML aplicados a cinco datasets de classificação binária, avaliando o desempenho dos modelos. A métrica AUC foi utilizada como principal critério de comparação. Sem ensembles, AutoGluon e MLJAR alternaram as melhores colocações. Com ensembles, técnicas que combinam múltiplos modelos para reduzir erros e melhorar a precisão, o AutoGluon venceu em todos os datasets.*

1. Introdução

Ferramentas de AutoML surgem para democratizar o uso do ML, automatizando etapas complexas e possibilitando que profissionais de diversas áreas apliquem técnicas avançadas sem profundo conhecimento técnico [Jidney et al. 2023]. Diante da variedade de frameworks disponíveis, torna-se essencial compará-los não apenas em termos de desempenho, mas também quanto à clareza dos resultados produzidos. Nesse contexto, a pesquisa se propõe a investigar e comparar diferentes frameworks de AutoML, considerando seu desempenho em cenários tabulares desafiadores.

2. Referencial Teórico

O desenvolvimento de ferramentas de Automated Machine Learning (AutoML) busca não apenas automatizar etapas do aprendizado de máquina, mas também maximizar o desempenho dos modelos gerados [Shen et al. 2018]. O problema de seleção de algoritmos[Rice 1976], evoluiu para incluir ajustes automáticos de hiperparâmetros, engenharia de atributos e geração de ensembles, formando pipelines completos orientados a resultados [Baratchi et al. 2024].

Estudos recentes, como de [Metin and Bilgin 2024], [Villarreal-Torres et al. 2024], também destacam que a escolha do framework impacta diretamente a capacidade de generalização do modelo e a estabilidade dos resultados, reforçando a necessidade de avaliações sistemáticas em cenários variados.

3. Metodologia

Este trabalho realiza uma avaliação de frameworks de AutoML utilizando datasets disponíveis em repositórios online, sem ajustes manuais prévios, delegando aos próprios frameworks todas as etapas de limpeza e adequação dos dados. A seleção das ferramentas considerou critérios como: ser de código aberto, possuir atualizações recentes, apresentar bons resultados em estudos recentes e dispor de documentação clara e acessível.

Os experimentos foram conduzidos em máquina com processador Intel® Core i5-13400F, 32 GB RAM, GPU NVIDIA GeForce RTX 4060 Ti 8 GB, SSD de 2TB, Windows 10 e Python 3.11.6.

A comparação entre os resultados dos frameworks foi realizada com base na métrica AUC (Área sob a Curva ROC). Os experimentos foram conduzidos em duas etapas, na primeira sem a possibilidade dos frameworks realizarem ensemble e na segunda com essa possibilidade. Ensemble é considerado um método sólido em aprendizado de máquina, pois combina vários modelos, como redes neurais artificiais ou outros algoritmos, com o objetivo de aumentar a exatidão das previsões. Em geral, um conjunto de modelos tende a gerar resultados mais consistentes e eficientes do que o uso de um único modelo individual [Shahab Hosseini and Sabri 2023].

4. Frameworks Selecionados

Foram analisados cinco frameworks de AutoML, selecionados com base em atualizações recentes, acesso ao melhor modelo e resultados em estudos na área, sendo eles: (1) AutoGluon, versão 1.3.1, de 22 de Maio de 2025; (2) H2O AutoML, versão 3.46.0.7, de 27 de Março de 2025; (3) MLJAR, versão 1.1.17, de 01 de abril de 2025; (4) EvalML, versão 0.84, de 08 de junho de 2024; (5) PyCaret, versão 3.3.2, de 27 de abril de 2024.

5. Datasets

As bases de dados utilizadas são compostas de dados tabulares foram coletadas de repositórios online com variados números de amostras e atributos, conforme Tabela 1.

Tabela 1. Descrição dos datasets utilizados

Dataset	Descrição	Número de Amostras	Número de Atributos
Adult Income Dataset	Predição de renda acima ou abaixo de $50k/ano com base em atributos demográficos.	48842	14
Airline Passenger Satisfaction	Predição da satisfação de passageiros de companhias aéreas.	129880	23
Bank Marketing Dataset	Previsão de aceitação de ofertas em campanhas de marketing.	45211	16
Credit Card Fraud Detection	Identificação de transações fraudulentas.	284807	30
Myocardial infarction complications	Previsão se um paciente tem ou não edema pulmonar.	1700	124

Tabela 2. AUC por dataset e framework (modelo entre parênteses)

Dataset \ Framework	AutoGluon	EvalML	MLJAR	H2O	PyCaret
Adult	0.931263 *(xgboost)*	0.929815 *(xgboost)*	0.929724 *(xgboost)*	0.929399 *(gbm)*	0.925945 *(lightgbm)*
Airline	0.995806 *(lightgbm)*	0.991617 *(lightgbm)*	0.995241 *(xgboost)*	0.995563 *(gbm)*	0.993917 *(lightgbm)*
Bank Marketing	0.930373 *(mlp)*	0.931272 *(xgboost)*	0.929773 *(xgboost)*	0.933877 *(gbm)*	0.931850 *(lightgbm)*
Credit Card	0.956342 *(mlp)*	0.970430 *(extra_trees)*	0.975703 *(xgboost)*	0.957975 *(gbm)*	0.929284 *(lr)*
Myocardial	0.745840 *(catboost)*	0.803774 *(extra_trees)*	0.805651 *(rf)*	0.767045 *(gbm)*	0.705205 *(gbm)*

6. Experimentos

Na primeira etapa da análise, conforme Tabela 2, observou-se que, considerando os cinco datasets e a execução sem uso de ensembles, os frameworks AutoGluon e MLJAR obtiveram o melhor resultado em mais datasets, liderando em dois datasets cada. Entre os modelos que garantiram o primeiro lugar para seus respectivos frameworks, o XGBoost foi o mais recorrente, aparecendo como vencedor em dois datasets. Ao analisar a frequência geral de seleção de modelos ao longo de todos os datasets, o XGBoost destacou-se novamente como o mais utilizado.

Tabela 3. AUC por dataset e framework com ensembles (modelo na linha inferior)

Dataset \ Framework	AutoGluon	EvalML	MLJAR	H2O	PyCaret
Adult	0.932543 *(ensemble)*	0.929815 *(xgboost)*	0.932020 *(ensemble)*	0.930206 *(ensemble)*	0.925945 *(lightgbm)*
Airline	0.996293 *(ensemble)*	0.991617 *(lightgbm)*	0.996103 *(ensemble)*	0.995808 *(ensemble)*	0.993917 *(lightgbm)*
Bank Marketing	0.938223 *(ensemble)*	0.930764 *(lightgbm)*	0.937217 *(ensemble)*	0.936767 *(ensemble)*	0.931850 *(lightgbm)*
Credit Card	0.982522 *(ensemble)*	0.972007 *(ensemble)*	0.979136 *(ensemble)*	0.959682 *(ensemble)*	0.929284 *(lr)*
Myocardial	0.832995 *(ensemble)*	0.803774 *(extra_trees)*	0.825081 *(ensemble)*	0.804383 *(ensemble)*	0.705205 *(gbm)*

Na etapa em que os frameworks tiveram a possibilidade de realizar ensemble de modelos, disponível na Tabela 3, o AutoGluon se destacou amplamente, conquistando a primeira colocação em todos os cinco datasets analisados. Em todos esses casos, o modelo vencedor foi justamente um ensemble, o que reforça a eficiência dessa estratégia em melhorar o desempenho preditivo. Ao observar a frequência geral de modelos selecionados ao longo de todos os datasets, o ensemble também liderou com ampla vantagem, aparecendo 16 vezes.

7. Conclusão

Os resultados obtidos neste estudo evidenciam que a escolha do framework de AutoML influencia significativamente o desempenho final dos modelos de classificação binária.

Na execução sem ensembles, observou-se um equilíbrio maior entre os frameworks, com AutoGluon e MLJAR alternando as melhores colocações e utilizando predominantemente algoritmos como XGBoost, GBM e LightGBM. Por outro lado, quando o uso de ensembles foi permitido, o AutoGluon apresentou domínio absoluto, vencendo em todos os datasets e reforçando a capacidade dessa abordagem de combinar modelos para alcançar maior poder preditivo.

No contexto geral da avaliação, o AutoGluon demonstrou o melhor desempenho entre os frameworks avaliados, destacando-se especialmente pela eficiência no uso de ensembles para gerar melhores resultados. Além dessa capacidade, destaca-se também pelo fácil acesso às informações detalhadas do processo de treinamento, permitindo uma análise mais completa e transparente dos resultados. Esse conjunto de características, aliado à facilidade de configuração e à documentação abrangente, reforça o potencial do AutoGluon como uma solução poderosa, versátil e amigável para aplicações que exigem alta performance.

Referências

Baratchi, M., Wang, C., Limmer, S., van Rijn, J. N., Hoos, H., Bäck, T., and Olhofer, M. (2024). Automated machine learning: past, present and future. *Artificial Intelligence Review*, 57(5):1–88.

Jidney, T. T., Biswas, A., Nasim, M. A. A., Hossain, I., Alam, M. J., Talukder, S., Hossain, M., and Ullah, M. A. (2023). Automl systems for medical imaging.

Metin, A. and Bilgin, T. T. (2024). Automated machine learning for fabric quality prediction: a comparative analysis. *PeerJ Computer Science*.

Rice, J. R. (1976). The algorithm selection problem. In *Advances in computers*, volume 15, pages 65–118. Elsevier.

Shahab Hosseini, Rashed Pourmirzaee, D. J. A. and Sabri, M. M. S. (2023). Prediction of ground vibration due to mine blasting in a surface lead–zinc mine using machine learning ensemble techniques. *Scientific Reports*.

Shen, Z., Zhang, Y., Wei, L., Zhao, H., and Yao, Q. (2018). Automated machine learning: From principles to practices. *arXiv preprint arXiv:1810.13306*.

Villarreal-Torres, H., Ángeles-Morales, J., Cano-Mejía, J., Mejía-Murillo, C., Flores-Reyes, G., Cruz-Cruz, O., Urcia-Quispe, M., Palomino-Márquez, M., Solar-Jara, M. Á., and Escobedo-Zarzosa, R. (2024). Comparative analysis of performance of automl algorithms: Classification model of payment arrears in students of a private university. *EAI Endorsed Transactions on Scalable Information Systems*, 11(4).