Smart Model Elimination for Efficient Automated Machine Learning

Eric Su Zhang[*]
St. Mark's School of Texas
10600 Preston Rd, Dallas, TX 75230
ericspring08@gmail.com

Benjamin Standefer
St. Mark's School of Texas
10600 Preston Rd, Dallas, TX 75230
bjmstandefer@gmail.com

Abstract

Automated Machine Learning or AutoML has emerged as a popular field of research. Many of these frameworks optimize a form of model selection, however, they all require every model to be run and evaluated. We propose a novel framework that automatically eliminates models that are unlikely to be performant by training a Boosting Model on hundreds of kaggle datasets. Out of the 30 model options, our framework can predict a top five model within its top eight 80% of the time.

1 Introduction

Machine learning (ML) is a type of artificial intelligence (AI) that uses a variety of algorithms to interpret data and extrapolate from it, making predictions based on observed trends. Individual models are trained on datasets to become smarter so as to make predictions based on new data. The development and optimization of these models is a time-consuming process involving statistical nuances and complex learning strategies. This has led to the emergence of Automated Machine Learning (AutoML) frameworks and research.

One sector of the global industry that has the most potential for beneficial integration of AutoML frameworks is healthcare, particularly in developing countries. These countries have severe shortages of healthcare workers and limited tools for diagnosis. For example, Africa has 2.3 healthcare workers per 1000 individuals, while the Americas have 24.8 healthcare workers per 1000. The World Health Organization (WHO) emphasized that this deficit is growing every year and will likely reach 18 million personnel by 2030. The use of AI in healthcare has been varied. Medical AI's typically automate repetitive tasks, making time consumption a primary concern. This coupled with a lack of adequate resources makes designing faster diagnostic systems for developing countries' medical sectors a pivotal issue. Most AutoML frameworks implement some form of model selection where a pool of models are filtered until a final model is selected. However, these frameworks require that every model to be run to filter out. In theory, this means that much energy is spent training models that are unlikely to be selected. We propose SMEML, a novel model selection algorithm that automatically eliminates models that it believes will not be performant.

1.1 Survey

Factor	**H2O AutoML**	**TPOT**	**MLJAR**
Ease of Use	Easy-to-use GUI and API	Python library, requires coding experience	Very user-fr

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

Factor	**H2O AutoML**	**TPOT**	**MLJAR**
Supported Algorithms	Linear, Tree, DNN, Clustering, etc.	Genetic Programming	Linear, Tree
Feature Engineering	Basic	Basic	Advanced -
Time Management	Time Constraints are Settable	Generations and population size	Time constr
Ensemble Methods	Yes (stacking + blending)	Yes (stacking)	Yes (stackin
Hyperparameter Tuning	Automated	Genetic algorithm	Automated
Model Interpretability Basic (feature importance)	SHAP, LIME, PDPs, MOJO	POJO, etc.	Basic
Scalability	Good (distributed computing)	Limited	Good
Customization	High	Moderate	High

2 Methodology

2.1 Data Collection

2.2 Model Elimination

3 Experiments

4 Discussion

5 Conclusion

Acknowledgments and Disclosure of Funding

References