International Journal of Public Health
HINTS AND KINKS
published: 13 April 2021
doi: 10.3389/ijph.2021.614296

Leveraging Automated Machine Learning for the Analysis of Global Public Health Data: A Case Study in Malaria

Elisabetta Manduchi[1,2]* and *Jason H. Moore*[1,2]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, United States, [2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

Keywords: automated machine learning, malaria, public health resources, genetic programming, predictive modeling

Edited by:
Olaf von dem Knesebeck, University Medical Center Hamburg-Eppendorf, Germany

****Correspondence:***
Elisabetta Manduchi
manduchi@pennmedicine.upenn.edu

Received: *05 October 2020*
Accepted: *17 March 2021*
Published: *13 April 2021*

Citation:
Manduchi E and Moore JH (2021) Leveraging Automated Machine Learning for the Analysis of Global Public Health Data: A Case Study in Malaria. Int J Public Health 66:614296. doi: 10.3389/ijph.2021.614296

The traditional approach to modeling the determinants of an outcome of interest in epidemiology consists of employing statistical methodologies, typically regression-based analyses. An alternative, which is becoming increasingly appreciated in the field, is offered by machine learning (ML), i.e., algorithmic based strategies which take advantage of the increased computational power over the past decade. In supervised ML, a *training* dataset is used to learn a predictive model for a *target* (the ML term for the outcome, i.e., the dependent variable) from specified *features* (the ML term for the explanatory, i.e., the independent, variables). The predictive performance of the resulting model is then assessed on a hold-out *testing* dataset. One of the main appeals of ML is that it offers more flexibility in terms of the assumptions about the variables and their relationships.

ML approaches yielded improved Risk of Sepsis scores [1] and models for groundwater nitrate exposure [2], to cite a few examples. However, setting up a suitable ML pipeline for a specific task involves several steps and decisions, including data pre-processing, feature selection, feature engineering, selection of ML algorithm(s), and tuning of the algorithm hyperparameters [3]. Optimizing ML-based pipelines with respect to predictive performance is therefore labor intensive and requires considerable domain expertize. An exciting development in the field consists of methods which assist (potentially non-expert) users in the design and optimization of ML pipelines. These methods are termed automated machine learning (AutoML) [4]. The Tree-based Pipeline Optimization Tool (TPOT) [5, 6] is an AutoML which employs genetic programming [7] to explore pipelines consisting of combinations of feature selection, feature transformation, and classification or regression steps and recommends the pipeline with the best performance. Good practice is to then assess the predictive performance of this pipeline on a separate hold-out testing dataset. TPOT offers a choice of several metrics to assess pipeline performance. In the application illustrated below we chose accuracy, the proportion of correctly classified individuals.

TPOT has been successfully used in biomedical applications including genetics [5], metabolomics [8, 9], transcriptomics [10, 11], and toxicogenomics [10]. Here we describe an application to infectious disease epidemiology leveraging data from ClinEpiDB, a resource aimed at advancing global public health by facilitating the exploration and analysis of epidemiological studies [12]. We selected the India International Centers of Excellence in Malaria Research (ICEMR) cross-sectional study of malaria in different transmission settings [13]. In the traditional regression-based analyses on this dataset reported in [14], the independent variables were age bin (<5, 5–14, 15+), gender, a history of travel in the two weeks preceding the survey visit, a history of malaria in the past year, antimalarial use in the two

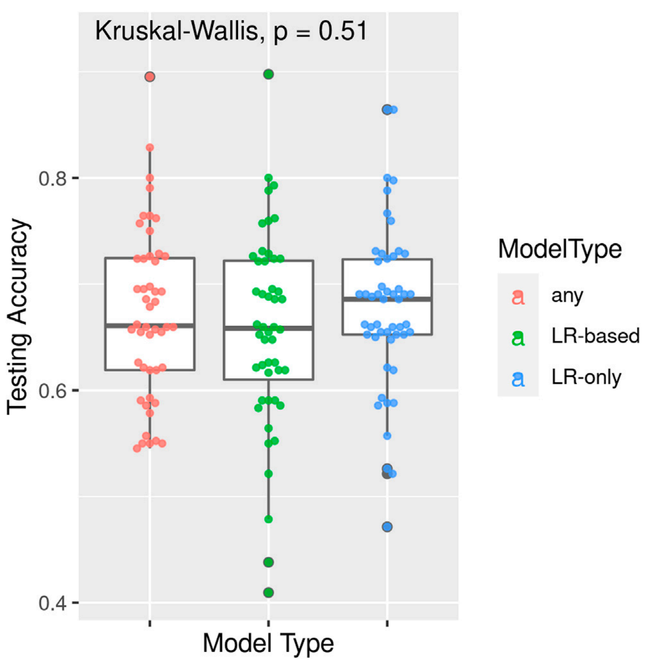

FIGURE 1 | Results of 50 TPOT runs for each of model types (i) a single regularized logistic regression step, (ii) a combination of feature selection, feature transformation, and regularized logistic regression steps, and (iii) a combination of feature selection, feature transformation and classification steps. Each point corresponds to a run and its *y*-coordinate indicates the accuracy of the pipeline optimized in that run over the hold-out testing dataset. Data are from the ICEMR cross-sectional study of malaria, India, 2012–2014.

weeks preceding the visit, reported use of repellent, and whether the visit occurred during the rainy season. There were five outcomes of interest (malaria detected by PCR, or by microscopy, submicroscopic/symptomatic/asymptomatic malaria) and analyses were stratified by site in India (Chennai/Nadiad,/Rourkela) and by *Plasmodium* species (*falciparum*/*vivax*/any), for a total of 45 analyses. We analyzed the same stratified combinations of targets and features using TPOT. As an illustration, we describe our TPOT results for one of these combinations, namely the analysis on individuals from Nadiad where the (binary) target is whether the individual had malaria (any species) detected by microscopy (58 cases, 793 controls). None of the features listed above was identified as significantly associated to the outcome in the regression analyses reported in [14]. We used TPOT to explore three increasingly complex types of pipelines: i) a single regularized logistic regression (LR) step, or ii) a combination of feature selection, feature transformation, and LR steps, or iii) a combination of feature selection, feature transformation and classification steps. For each of these three types, we ran TPOT 50 times with different random splits of the input data into training (75%) and hold-out testing (25%) portions. Moreover, to mitigate the effect of the high imbalance between number of cases and controls, in each run we randomly undersampled the controls to equal the number of cases prior to the random split. **Figure 1** summarizes the results, where each point represents one of the 50 runs of TPOT exploring pipelines of the type indicated by its color. On the *y*-axis the accuracy on the testing set of the TPOT-optimized pipeline for that run is indicated. The Kruskal-Wallis test did not detect a significant difference in the three distributions. For this dataset, simple pipelines of type (i) achieve good microscopic malaria prediction accuracy on average (mean = 0.68), when the regularized LR hyperparameters are tuned. Models of type (i) generalize those used in [14], and are easily interpretable. The best accuracy across these 50 TPOT runs is 0.86. On the other hand, the best accuracies across type (ii) and (iii) runs are ~90%. There is a tradeoff between interpretability and complexity. **Figure 2** depicts the architecture of the best pipeline of type (iii). This has a higher accuracy than the best LR pipeline but is quite complex and it

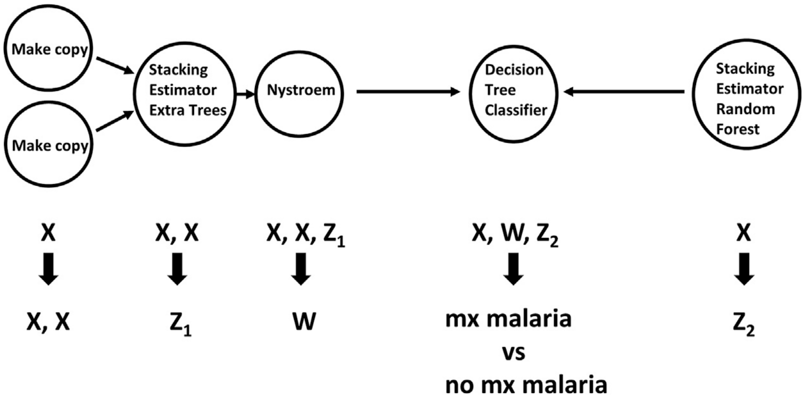

FIGURE 2 | Workflow of feature selection, feature transformation and classification steps for the pipeline of type (iii) with the best accuracy. Arrows indicate step outputs which are fed as inputs of subsequent steps. When exploring pipeline space, currently the Tree-based Pipeline Optimization Tool can choose among five feature selection, 14 feature transformation and 15 classification algorithms. The node labels in this graph indicate the algorithms selected in this pipeline. Nystroem is a feature transformation algorithm; Extra Trees, Decision Tree and Random Forest are classifiers. The Stacking Estimator adds to its input features the results of applying the indicated classifier to those features. The Tree-based Pipeline Optimization Tool tuned hyperparameter values for each step are not shown for visual simplicity. Data are from the ICEMR cross-sectional study of malaria, India, 2012–2014.

would have unlikely been discovered without an AutoML approach. The most relevant features in this pipeline, in terms of driving the predictive model, were gender and antimalarial use, based on permutation importance, which is a standard ML method to aid in model interpretability, described at https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html#eli5-permutation-importance.

This example underscores the utility of AutoML approaches in epidemiology, especially those offering to non-expert users the ability to specify the type of pipelines to explore, from very simple to very complex, at the same time leaving the heavy lifting to the AutoML. We add that the most recent extension of TPOT enables covariate adjustments [10] which in some epidemiology settings is crucial. Embedding AutoML tools within epidemiology platforms like ClinEpiDB would empower users to directly perform sophisticated analyses, accelerating the benefits derived from these public health resources.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://clinepidb.org/ce/app/record/dataset/DS_a5c969d5fa.

AUTHOR CONTRIBUTIONS

JM conceived and supervised the project. EM analyzed the data and drafted the manuscript. Both authors reviewed and edited the manuscript. Both authors read and approved the final manuscript.

FUNDING

National Institutes of Health grant AI116794. The funding body had no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

We thank Weixuan Fu for clarifications on TPOT pipeline representations.

REFERENCES

1. Delahanty, RJ, Alvarez, J, Flynn, LM, Sherwin, RL, and Jones, SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann Emerg Med* (2019) 73(4):334–44. doi:10.1016/j.annemergmed.2018.11.036
2. Messier, KP, Wheeler, DC, Flory, AR, Jones, RR, Patel, D, Nolan, BT, et al. Modeling groundwater nitrate exposure in private wells of North Carolina for the agricultural health study. *Sci Total Environ* (2019) 655:512–9. doi:10.1016/j.scitotenv.2018.11.022
3. Wiemken, TL, and Kelley, RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health* (2020) 41(1):21–36. doi:10.1146/annurev-publhealth-040119-094437
4. F Hutter, L Kotthoff, and J Vanschoren, editors. *Automated machine learning: methods, systems, challenges*. Springer International Publishing (2019) Available from: https://www.springer.com/gp/book/9783030053178. (Accessed March 31, 2021)
5. Olson, RS, Urbanowicz, RJ, Andrews, PC, Lavender, NA, Kidd, LC, and Moore, JH. Automating biomedical data science through tree-based pipeline optimization. In: G Squillero and P Burelli, editors. *Applications of evolutionary computation*. Cham: Springer International Publishing (2016) p. 123–37. (Lecture Notes in Computer Science).
6. Olson, RS, and Moore, JH. TPOT: a tree-based pipeline optimization tool for automating machine learning. In: F Hutter, L Kotthoff, and J Vanschoren, editors. *Automated machine learning: methods, systems, challenges [internet]*. Cham: Springer International Publishing (2019) doi:10.1007/978-3-030-05318-5_8
7. Koza, JR. *Genetic programming: on the programming of computers by means of natural selection*. MIT Press (1992) p. 858.
8. Orlenko, A, Moore, JH, Orzechowski, P, Olson, RS, Cairns, J, Caraballo, PJ, et al. World Scientific. Considerations for automated machine learning in clinical metabolic profiling: altered homocysteine plasma concentration associated with metformin exposure, *Biocomputing* (2017) 460–71. doi:10.1142/9789813235533_0042 Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813235533_0042 (Accessed March 31, 2021).
9. Orlenko, A, Kofink, D, Lyytikäinen, L-P, Nikus, K, Mishra, P, Kuukasjärvi, P, et al. Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. *Bioinformatics* (2020) 36(6):1772–8. doi:10.1093/bioinformatics/btz796
10. Manduchi, E, Fu, W, Romano, JD, Ruberto, S, and Moore, JH. Embedding covariate adjustments in tree-based automated machine learning for biomedical big data analyses. *BMC Bioinformatics [Internet* (2020) Available from: https://link.springer.com/epdf/10.1186/s12859-020-03755-4 (Accessed March 31, 2021).
11. Le, TT, Fu, W, and Moore, JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Kelso J Editor Bioinformatics* (2020) 36(1):250–6. doi:10.1093/bioinformatics/btz470
12. Ruhamyankaka, E, Brunk, BP, Dorsey, G, Harb, OS, Helb, DA, Judkins, J, et al. ClinEpiDB: an open-access clinical epidemiology database resource encouraging online exploration of complex studies. *Gates Open Res* (2020) 3:1661. doi:10.12688/gatesopenres.13087.2
13. Carlton, J, Eapen, A, Sharma, SK, Srivastava, HC, Singh, OP, and Valecha, N. India ICEMR Cross-sectional study. *ClinEpiDB* (2019) 7. 2019 Available from: https://clinepidb.org/ce/app/record/dataset/DS_a5c969d5fa (Accessed March 31, 2021).
14. van Eijk, AM, Sutton, PL, Ramanathapuram, L, Sullivan, SA, Kanagaraj, D, Priya, GSL, et al. The burden of submicroscopic and asymptomatic malaria in India revealed from epidemiology studies at three varied transmission sites in India. *Sci Rep* (2019) 9(1):17095. doi:10.1038/s41598-019-53386-w