Improving Efficiency in Analysis of Real-World Data With an Automated Machine Learning Tool

Poster number: MSR63

Scan QR

Ying Zhang[1], Wei-Hsuan "Jenny" Lo-Ciganic,[2,3] Handing Xie,[1] Ravi Iyer,[1] Dennis Snyder,[1] Pete Lineman,[1] Marc Tian[1]

[1]Teva Branded Pharmaceutical Products R&D, Inc., West Chester, PA, United States; [2]University of Pittsburgh, Pittsburgh, PA, United States; [3]North Florida/South Georgia Veterans Health System, Geriatric Research Education and Clinical Center, Gainesville, FL, United States

Objective: To evaluate the validity and efficiency of an automated machine-learning (ML) tool compared to traditional ML approaches in a real-world data (RWD) analysis

Background

- Real world data (RWD) are becoming increasingly important sources to generate evidences that can guide decision-making in clinical development and in the life cycle management of medical products.
- Machine learning (ML) is increasingly being employed to extract novel clinical insights from RWD (eg, predictions of mortality,[1] risk of readmission,[2] and medication adherence[3]).
- ML requires sophisticated analytical skills and can be time-consuming to optimize model performance.
- Having automated ML analytical tool can be a promising approach to improve the efficiency of model development and greatly accelerate the insights and evidence generation through RWD.

Methods

- AutoML is a point-and-click ML tool on a cloud-based platform developed by Databricks Inc., utilizing big data computing resources and automatizing the ML process,[4] such as:
  - Identifying feature types and feature engineering
  - Fine-tuning hyperparameters
  - Training and validating multiple ML algorithms in parallel
- We leveraged a case RWD analysis to evaluate the performance of AutoML compared to the ML approaches with hand-coding:
  - **Case study objective:** Predict the treatment instability and identify important predictors in patients with schizophrenia initiating oral antipsychotics using Merative™ MarketScan® claims databases
  - **Cohort criteria:** Patients with schizophrenia who had ≥1 oral antipsychotic claims (2013 –2021); continuous enrollment in the 1-year baseline and 6-months follow-up
  - **Baseline features:** demographic variables, medical diagnosis and procedure claims, pharmacy claims, healthcare resource utilizations such as emergency department visit and hospitalization

Key Results

Figure 1. ML Running Time (hour)

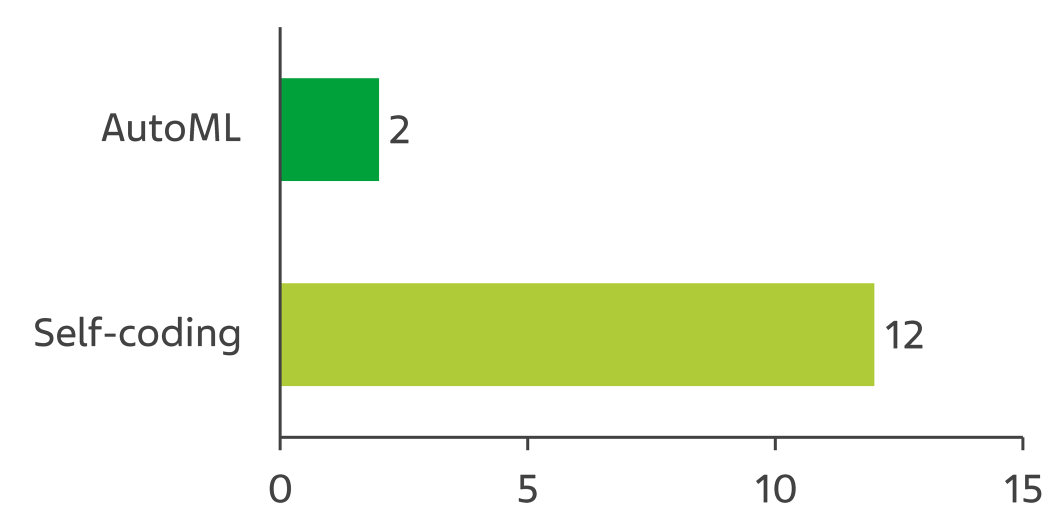


Table 1. Performance Metrics for AutoML and Traditional ML Analysis

		AUC	Accuracy	Precision	Recall	F1 score
AutoML	XGBoost	0.64	0.63	0.86	0.64	0.74
	Random forest	0.58	0.68	0.83	0.76	0.79
	Elastic Net	0.58	0.61	0.84	0.64	0.72
Self-coding	XGBoost	0.61	0.65	0.84	0.70	0.76
	Random forest	0.61	0.67	0.84	0.73	0.78
	Elastic Net	0.64	0.67	0.85	0.73	0.78

- **ML models**: Elastic net, random forest, and XGBoost (XGB) were trained and tested by AutoML and self-coding ML approach using Python 3.8.10
- Processes for applying AutoML

Prepare analysis dataset
- Save the analysis dataset with finalized cohort into Databrick
- Specify the feature types (Numerical/Categorical) in Spark system

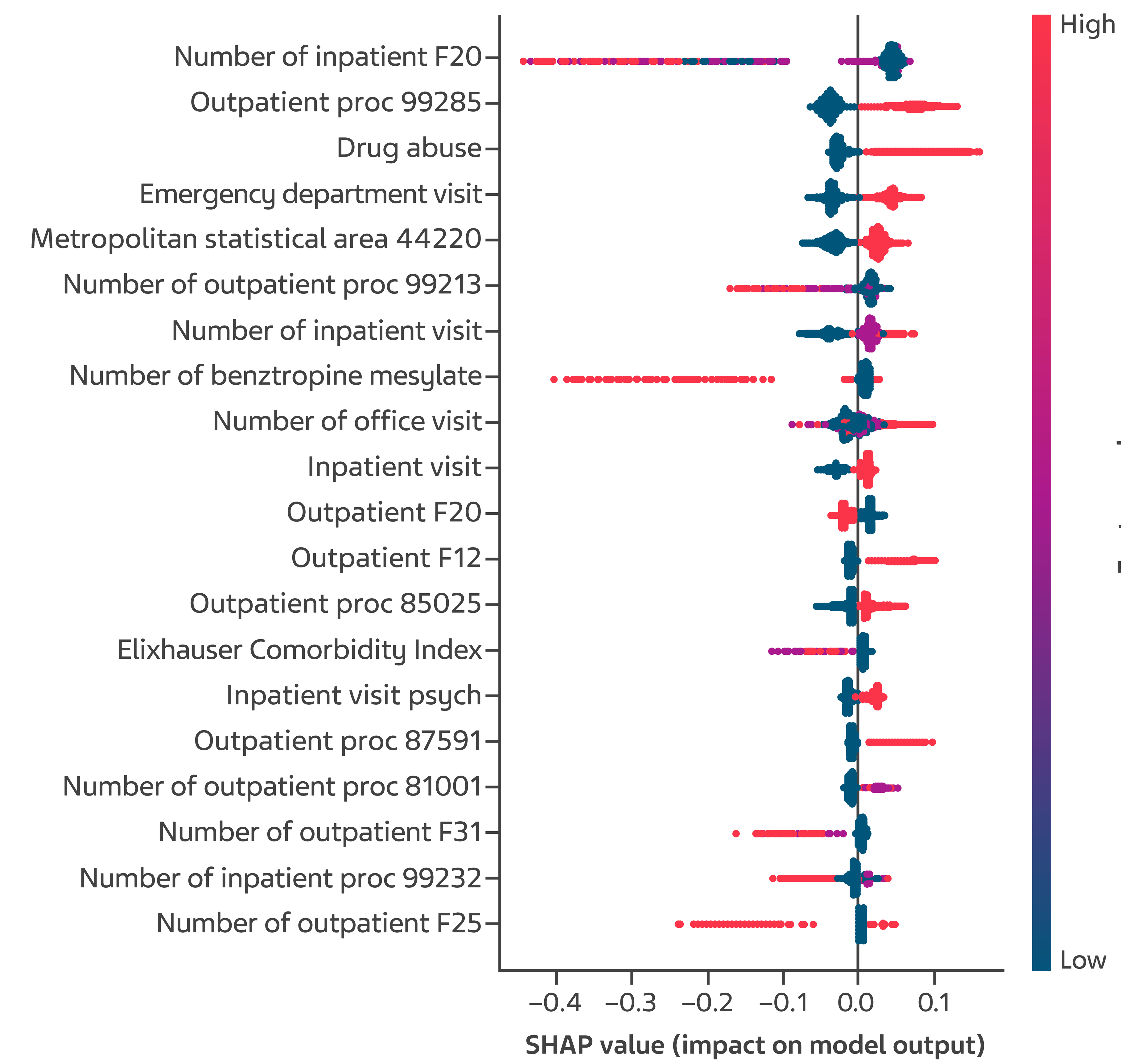
Specify key elements in ML analysis
- Select the cloud computing cluster with AutoML capacity
- Define the type of ML models: classification/prediction and main evaluation metrics
- Choose ML algorithms from random forest, XGB, elastic net, regression

Interpret the ML results
- Use the outputs panel to evaluate the performance of different ML models
- Use auto-generated Python notebooks to filter important features

Additional Results

- The analysis cohort included 4671 adults; 80.9% of patients had treatment instability as the outcome with 1549 claims-based features were included in the ML analysis
- The AutoML only required 16% of the computational time (2 vs 12 hours) compared with using self-coding ML approach (Figure 1) with similar results (Table 1)
  - **Best-performing model using AutoML:** XGBoost (AUC: 0.64 vs. 0.58 using other methods) with high precision (0.86)
  - Elastic net using self-coding ML approach yielded similar prediction performance (AUC: 0.64) with high precision (0.85)
- AutoML identifies key predictive predictors for treatment instability, such as number of outpatient, diagnoses of schizophrenia and emergency department visits (Figure 2).
- Key predictors were largely overlapping between AutoML and self-coding approaches.

Figure 2. SHAP Summary Plot Using XGB Model in AutoML



Conclusion

- Our case study showed the automated ML tool has the potential to democratize and augment ML applications in RWD analysis
- With transparent source codes and results reports, it can accelerate further model optimization and improve efficiency

Disclosures

This study was supported by funding from Teva Branded Pharmaceutical Products R&D, Inc. Jenny Lo-Ciganic has received consulting fees and/or honoraria from Teva Pharmaceuticals. Ying Zhang, Handing Xie, Ravi Iyer, Dennis Snyder, Pete Lineman, and Marc Tian are employees and stockholders of Teva Pharmaceuticals.

Abbreviations

AUC = area under curve, ML = machine learning, RWD = real-world data, PROC = procedure, SHAP = SHapley Additive exPlanations

References

1. Seki T, et al. *PLoS One*. 2021;16:e0246640.
2. Mohanty SD, et al. *Patterns (NY)*. 2021;3:100395.
3. Lo-Ciganic, et al. *Med Care*. 2015; 53:720–728.
4. *Democratization of Advanced Analytics: Unveiling significance of AutoML in Databricks* https://www.linkedin.com/pulse/democratization-advanced-analytics-unveiling-automl-sreepriya-mohanan/?trackingId=NN13nmH2QcWLNUd60Xo%2FzQ%3D%3D

Presented at the Annual Meeting of the International Society for Pharmacoeconomics and Outcomes Research; May 5–8, 2024; Atlanta, GA.