TuftsMedicine
Tufts Medical Center

NEWDIGS

Outline of Machine Learning & Statistical Methods

LEAPS Methods Innovation Team

Case Study Background

For the initial case study, develop a predictive model(s) that can improve decision-making for all stakeholders related to immune checkpoint inhibitor use in patients with advanced NSCLC.

Objectives

Assess feasibility of using federated (machine) learning methods, leveraging diverse data types (e.g., EHR, claims, social determinants of health, biologic, clinical trials, patient-generated, etc.) to:

- Identify signals, generate hypotheses about clinically meaningful sub-populations
- Define next step in corroborating/validating promising hypotheses
- Reduce bias in algorithm development through the use of diverse data sets
- Establish federated learning environment (technology enablers, cross-functional expertise, governance) that is scalable

Purpose of the Machine Learning & Statistical Methods Outline

- Discuss a list of machine learning and statistical methods that can be applied to a series of case studies within LEAPS
- Identify the right machine learning/statistical models to fit for the appropriate response variables with emphasis on full transparency
- Develop a framework for assessing and validating the statistical models

Application and Approach

The Machine Learning (ML) & Statistical Methods Outline (Outline) seeks to identify and characterize the strengths and limitations of available machine learning and statistical methods to be applied to the Advanced NSCLC Use Case specifically and more generally to other use cases as identified by the LEAPS team. In addition to capturing general details, the strengths and limitations of the ML and statistical methods are characterized and assessed across multiple objective dimensions.

Strengths & Limitations

- Level of transparency of model or methods (low/high)
    - Significance: why an important dimension?

- Ease of interpretability by end user(s), e.g., clinicians
    - Significance: why an important dimension?

- Flexibility to add in new/different models, in addition to the initial model selection(s), is important
    - Significance: why an important dimension?

- Performance boost over a meaningful benchmark
    - Significance: why an important dimension?

- Retrospective vs. prospective interventional validation
    - Significance: why an important dimension?

- Potential algorithmic bias against key protected characteristics, including likely cause(s), and options for correcting the bias if possible
    - Significance: why an important dimension?

Other factors for consideration in model selection

- Balancing using statistically appropriate models with models that are accepted and useful to the end user(s), e.g., clinicians, payers, and other key decision-makers
    - Significance: why an important dimension?

- Recommendations from data partners on model preferences based on prior experience
    - Significance: why an important dimension?

Appendix

Table 1: List of ML and Statistical models under consideration

Method/Model	Description	Unique Feature(s)	Strengths	Limitations
Logistic Regression	The most common statistical model to understand the relationship between a binary response variable and a list of predictors. The model assumes linear additive relationship between predictors and the response variable	Clear interpretation of relationship between response and predictor variables	Easy to fit Ease of interpretability by end user(s)	1) Only for binary response variable can handle only small number of predictors. 2) Linear additive assumption between predictors and response variables is not necessarily hold in reality 3) Prediction performance is not as good as the other models can handle complex relationship between predictor and response variables
Decision Tree	A graph that applies a series of binary decision rules for the predictors to predict the response variables	Model can be interpreted as a list of decision rule	Ease of interpretability by end user(s)	1) Can handle only small number of predictors 2) Maybe too simple to handle complex relationship in reality
Random forest	A classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to	Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.	Can handle complex relationship	1) Need relatively large sample size 2) Not easy to interpret the model

Method/Model	Description	Unique Feature(s)	Strengths	Limitations
	create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.			
LASSO	Logistic regression with implemented for the	The lasso procedure encourages simple, sparse models.	Easy to fit Ease of interpretability by end user(s)	1) Can handle moderate number of predictors 2) Maybe too simple to handle complex relationship in reality
Neural Network/Deep Learning	Mimics the behavior of the brain		Can handle very complicated relationship between	Difficult to training the model, need large sample size Difficult to interpret the model
XGBoost	Implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.			Overfitting Difficulties in model interpretation