Privacy Pitfalls in Tree-Based Ensemble Models Developed using Electronic Health Record Data: A Case Study

Karandeep Singh, MD, MMSc[1], Adharsh Murali[1], Michael Burns, MD, PhD[1], Brahmajee K. Nallamothu, MD, MPH[1], Akbar K. Waljee, MD, MSc[1], and Jeremy B. Sussman, MD, MSc[1]

[1] Michigan Integrated Center for Healthcare Analytics and Medical Prediction, University of Michigan, Ann Arbor, MI

Background. As part of the international shift towards open and reproducible science, the practice of sharing predictive models has become more common in biomedical research. This practice generally does not violate patient privacy standards because models derived from identified data do not contain any of the 18 personal identifiers that constitute protected health information (PHI) in the Health Insurance Portability and Accountability Act (HIPAA). Models are commonly fit on data containing PHI with the goal of learning useful representations of the data. But with the growing diversity of predictive modeling methods, some models could themselves contain PHI. For example, publicly sharing a k-nearest neighbors model derived from PHI-containing data would clearly violate HIPAA because the model contains a copy of the data. Tree-based ensemble models, including random forests and gradient-boosted decision trees, also could reveal aspects of the underlying data, including PHI. This problem is rarely discussed for tree-based ensemble models. In Leo Breiman's original implementation, random forest classification trees are grown until each leaf contains a single observation (or multiple identical observations). Unlike decision trees, random forests are not pruned. Thus, demographic information contained within tree-based ensemble models could be combined with public data (e.g., voter records) to re-identify patients. We explored the extent of this problem using recently published tree-based ensemble models.

Methods. We identified a recent publication in which random forest (RF) models and gradient-boosted decision tree (GBDT) models trained on electronic health record data (EHR) were made publicly available. The study used RF and GBDT models to predict two separate delirium outcomes in newly hospitalized patients (Wong A *et al.* 2018). The models are available at https://github.com/ayoung01/delirium. We searched the trees to identify the number of splits above the age of 89. The HIPAA Privacy rule considers all ages over 89 to be PHI, so de-identified health datasets are required to aggregate all individuals in this age range into a single group, which is typically accomplished by top-coding age at 89 years. For all splits at ages greater than 89, we traversed the trees in both directions from the age split (to the root and leaf nodes) to identify other splits based on variables that may help re-identify individuals, such as gender and race.

Results. From this study, we identified hundreds of trees containing splits at ages greater than 89 years and several trees containing concurrent information on gender and race (Table 1). We identified a split at age > 101.5, indicating that at least one patient was above this age in the source dataset. According to the 2010 U.S. Census data, only 20,000 Americans fit this description. The models were trained on data from San Francisco (CA), which narrows this to a much smaller number of individuals. Based on traversing a random forest tree, we also identified an individual with age > 90.5 who is Asian, has known psychosis, and is on treatment for hypertension. While many states have publicly searchable voter records by age, California does not, so re-identification was not a concern in this current paper.

Table 1. Information about individuals over 89 years of age from publicly available models trained on EHR data

Model type	Delirium outcome	# Trees	# Trees with splits with age > 89	# Splits with age > 89	# Splits with age > 89 *and* gender	# Splits with age > 89 *and* race
RF	Primary	1,025	296	351	4	38
RF	Secondary	1,025	663	1,032	12	77
GBDT	Primary	21,300	827	371	0	1
GBDT	Secondary	22,500	334	865	1	5

Conclusion. We identified publicly available tree-based ensemble models that could have contained protected health information. While the models could not be linked to voter records in this instance, the prospect that this could be done in other states should be cause for concern. Capping age, limiting tree depth, or creating tools that prevent HIPAA violations are potential mechanisms to address this and should be considered by the machine learning community.