

Lee Xiong An

Holmusk Take home Test
15-22 Dec 2021

Problem Statement

- I. Compare the real world efficacy of Drug A vs Drug B using survival analysis
- II. Take into account confounding factors affect treatment as data is EHR and imbalanced

Dataset

- Patient_characteristics.csv – contains the information about socio-demographics of patients, diagnosis, lab values, and other existing therapies patients are on
- Event_duration.csv – Contains information about if the event happened to a patient or if patient was censored and the durations for each case from the start of therapy

patient_id	0
Bleeding_event (1=event, 0=censored)	0
duration_in_years	0.019178
treatment_variable	Drug_A
sex	1
age	86
other_drugs_1	0
other_drugs_2	0
other_drugs_3	0
other_drugs_4	0
other_drugs_5	0
other_drugs_6	0
other_drugs_7	0
other_drugs_8	1
diagnosis_1	0
diagnosis_2	0
diagnosis_3	0
diagnosis_4	0
diagnosis_5	0
diagnosis_6	0
diagnosis_7	0
diagnosis_8	1
diagnosis_9	0
diagnosis_10	1
diagnosis_11	0
diagnosis_12	0
diagnosis_13	1
diagnosis_14	0
diagnosis_15	0
lab_1	0.974967
lab_2	358.572186
lab_3	NaN
lab_4	NaN
lab_5	5.49857
lab_6	17.060713
lab_7	1.077035
lab_8	67.781782
Diag_Score_1	5
Diag_Score_2	5

Approach for Problem

- I. Method to detect confounding using decision tree
- II. Obtaining balanced dataset using (I)
- III. Validating that balanced dataset has reduced confounding
- IV. Compare efficiency of Drug A & B using univariate KM on balanced dataset
- V. Limitations of method

Confounding factors 101 (for the unfamiliar)

- For EHR data, results may be confounded unlike that of a clinical trial where everything is controlled and balanced
- If sicker patients were given treatment with drug A, drug A would be associated with poorer outcomes and vice versa
 - Drug A is confounded with “sick”, hence analysis is likely invalid
- A remedy is to rebalance the dataset such that both drugs have similar profiles of people

Method to Detecting confounding

- What is a balanced dataset with respect to drug type?
 - A balanced dataset would have drug type indistinguishable
- Building a machine learning model with patient profile and predict drug type administered
 - If confounding is strong, the model will predict drug type well
 - If confounding is small, the model will predict poorly, i.e. patient profile not associated with drug type
 - Model evaluated by ROC AUC

ROC AUC (for unfamiliar)

- ROC AUC is a simple metric that is easy to understand, $P(X_1 > X_0)$ proof below
- 0.5 means model is random, 1 mean model is perfect, 0 means model is perfectly wrong

$$\text{TPR}(T) : T \mapsto y(x)$$

$$\text{FPR}(T) : T \mapsto x$$

$$A = \int_{x=0}^1 \text{TPR}(\text{FPR}^{-1}(x)) dx = \int_{-\infty}^{\infty} \text{TPR}(T) \text{FPR}'(T) dT = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(T' > T) f_1(T') f_0(T) dT' dT = P(X_1 > X_0)$$

where X_1 is the score for a positive instance and X_0 is the score for a negative instance, and f_0 and f_1 are probability densities as defined in previous section.

Results : Confounding Model

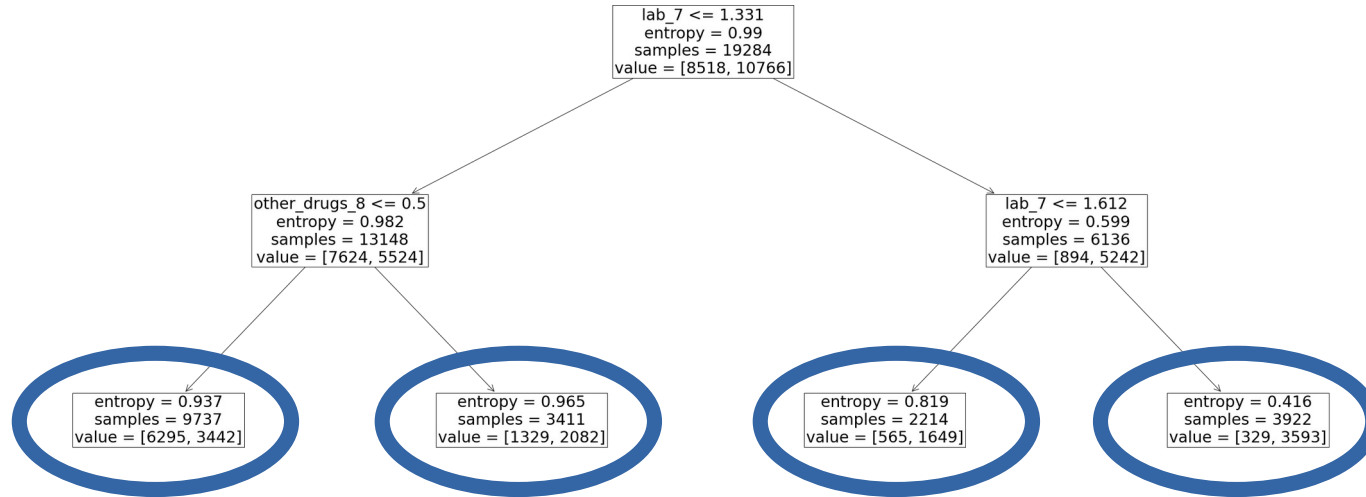
- DecisionTreeClassifier model used, maximum_depth = 6 and min samples in final nodes = 15
 - Decision Tree is simpl model to split population into subgroups
 - Explanatory variable : patient characteristic
 - Dependent Variable : Drug Type
 - 10 fold CV used
 - ROC AUC = 0.79
- Data is definitely confounded
 - Even the presence of null values give RO AUC 0.65

Obtaining balanced dataset

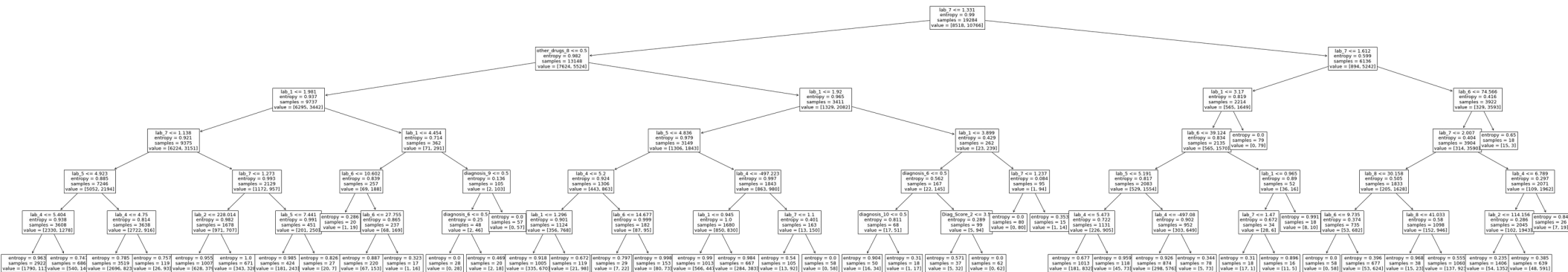
- Subsample patients based decision tree predicting drug type

Obtaining balanced dataset

- Subsample patients based decision tree predicting drug type from patient features
- Subsample balanced dataset from terminal node of decision tree



Tree generating balanced dataset



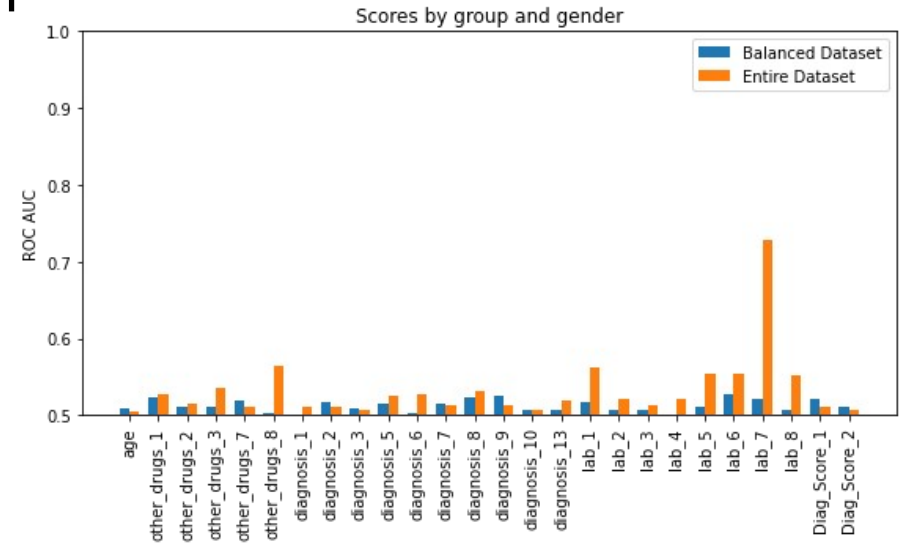
- DecisionTreeClassifier model used, maximum_depth = 6 and min samples in final nodes = 15
 - 10 fold CV ROC AUC = 0.79
- This tree is generated with 100% of data

Characteristics of balanced dataset

- Decision tree has profiled patients into 44 nodes of which 37 nodes have both patients with drug A and B
 - These 37 nodes make up 98% of the dataset
 - The other 7 nodes only have treatment=Drug B and dropped from balanced dataset
- A balanced dataset of 10K (53%) is sampled from these patients

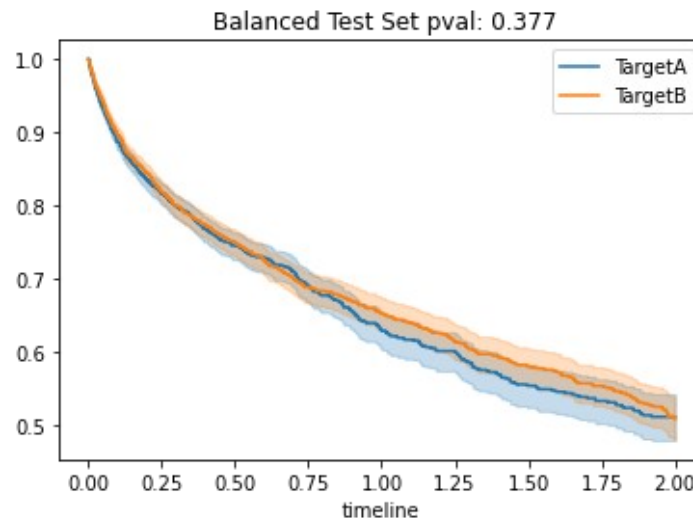
Metrics for Balanced Dataset

- Trained *univariate* decision tree with depth=3, min_leaf=30
 - Explanatory variable: X axis
 - Target variable: Drug A/B
 - 10 fold ROC AUC Score plotted
- Balanced dataset cannot be predicted by patient characteristics
 - Even using all patient information, model performs 0.55 ROC AUC



Compare Real World Efficacy of Drug A vs B

- Used a Log Rank test to compare efficacy
- Dataset is balanced hence test results hold for populations similar to the balanced dataset
- Plot shows KM curve and pval for Log Rank test



Limitations

- The balanced dataset may not be a true representative of the main population
 - Reweighing could be done to account for that
- Our balancing method removes the strongest confounders, there are also many weak confounders not accounted for
 - Drug type can still be predicted from patient characteristics in balanced dataset at ROC AUC = 0.55

Approach

- I. Method to detect confounding using decision tree
- II. Obtaining balanced dataset using (I)
- III. Validating that balanced dataset has reduced confounding
- IV. Compare efficiency of Drug A & B using univariate KM on balanced dataset
- V. Limitations of method