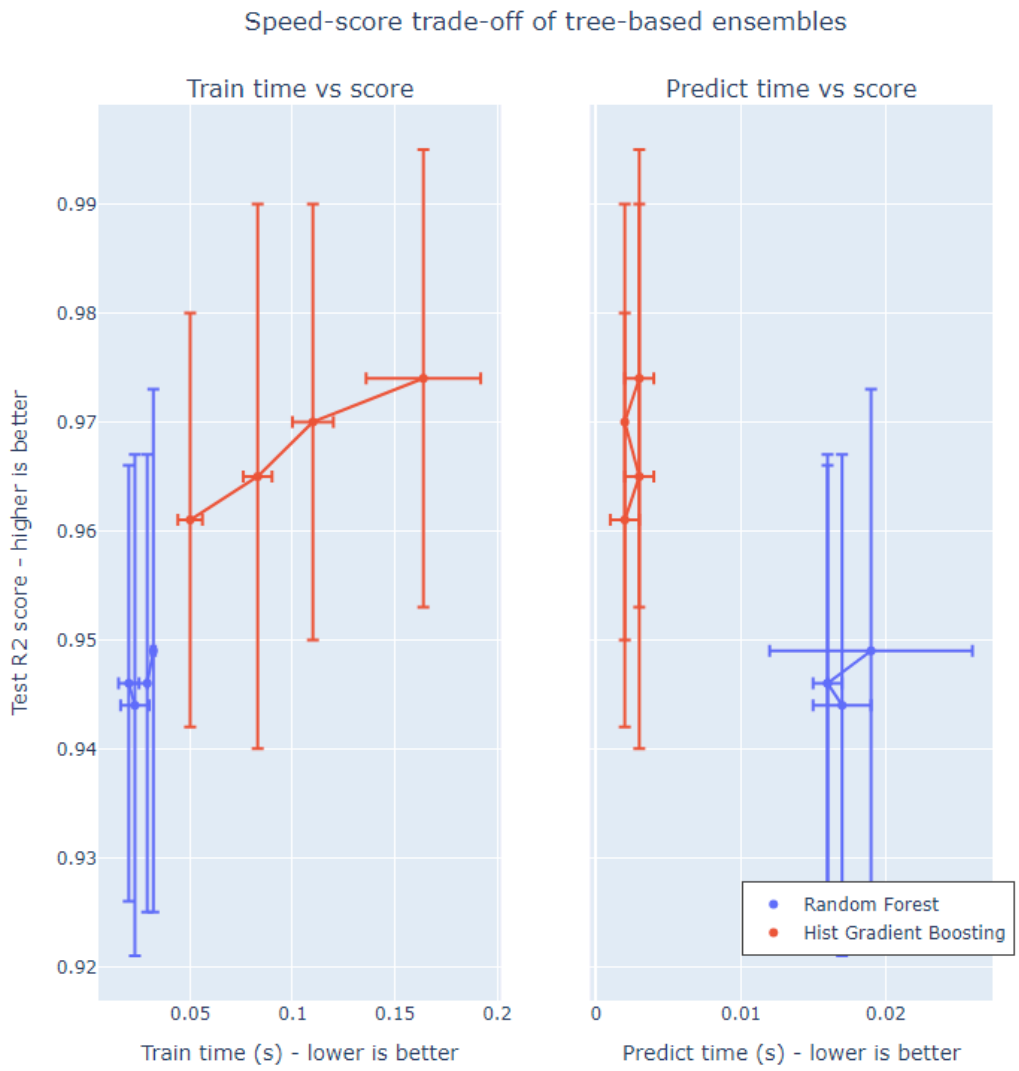
David Shubov

Problem 1:

a) Breast cancer dataset with the Random Forest Classifier on 10, 25, and 50 estimators

n_estimators	separator	average fold accuracy
10	gini	0.9508150908244062
10	entropy	0.9578636857630801
25	gini	0.9561092997981679
25	entropy	0.9631268436578171
50	gini	0.9613569321533924
50	entropy	0.9525850023288308

b) Random Forest vs Histogram Gradient Boosting



Problem 4: Utility

(5 points) Which route should we pick? Show your work. You may find it easier to convert km/hr to mins/km (how many minutes does it take to go 1km on each surface?)

Rocky:
$2km/h \ -> \ 60min \ / \ 2km/h \ = \ 30 \ min \ per \ km$

Sandy:
$3km/h \ -> \ 60min \ / \ 3km/h \ = \ 20 \ min \ per \ km$

Smooth:
$5km/h \ -> \ 60min \ / \ 5km/h \ = \ 12 \ min \ per \ km$

Route 1:
$(0.20 \ * \ 20 \ * \ 5) \ + \ (0.30 \ * \ 12 \ * \ 5) \ + \ (0.50 \ * \ 30 \ * \ 5) \ = \ 113 \ min$

Route 2:
$(0.40 \ * \ 20 \ * \ 7) \ + \ (0.20 \ * \ 12 \ * \ 7) \ + \ (0.40 \ * \ 30 \ * \ 7) \ = \ 156.8 \ min$

Route 3:
$(0.50 \ * \ 20 \ * \ 6) \ + \ (0.40 \ * \ 12 \ * \ 6) \ + \ (0.10 \ * \ 30 \ * \ 6) \ = \ 106.8 \ min$

We should pick route 3.

(5 points) Update your estimates for the travel time for each route. Now which route seems best?

Route 1 (initially 113 minutes)
Damaged wall: $0.30 \ * \ 15 \ = \ 4.5 \ min$
Intact wall: $0.70 \ * \ (-20) \ = \ -14 \ min$
New time: $113 \ + \ (4.5 \ - \ 14) \ = \ 103.5 \ min$

Route 3 (initially 106.8 minutes)
Damaged bridge: $0.60 \ * \ 40 \ = 24 \ min$
Intact bridge: $0.40 \ * \ 0 \ = \ 0 \ min$
New time: $106.8 \ + \ 24 \ = \ 130.8 \ min$

Route 1 is the new best.

(2 points) First: If the satellite said that route 2 was not rocky, how long would we expect it to take?

Only consider sandy and smooth: $(0.40 * 20 * 7) + (0.20 * 12 * 7) = 72.8\,min$

(3 points) Second: What's the probability that the satellite will tell us this?

Total probabilities must sum to 1, so since Route 2 has a 40% chance of being rocky then the remainder is 1 - 0.4 or 0.6 (60%) probability of the satellite telling us its not rocky.

(2 points) Third: If the satellite tells us route 2 is in fact rocky, what do we do? How long will that take?

We know that the terrain is definitely rocky so the expected time for rocky terrain is used: $30\,min/km * 7\,km = 210\,min$

(3 points) Last: given all of this, how long should we wait for the satellite?

So 60% chance of not rocky, with expected time being 72.8 min and 40% chance of rocky with expected time being 210 min.

Therefore, the expected wait time is: $(0.60 * 72.8) + (0.40 * 210) = 127.68\,min$