S2. Additional figures and table

2021-10-11

Contents

Posterior predictive distributions **2**

List of terms relevant for prediction **3**

Hyperparameters grid search **3**

Posterior predictive distributions

Figure 1 show the mixture of the predictive distribution of the probability of a positive match, grouped by labelling status. The posterior samples for each record were extracted and joined in a global distribution; on this distribution, density was computed on the logit scale and then logistic transformed for display on a $[0-1]$ scale. The light violet ridges show the distribution of the still unlabelled records.
For each iteration, the thresholds which define the lower and the upper range of 98% predictive interval (PrI) respectively for the positive and negative records are shown. The zone included between these two boundaries defines the uncertainty zone; records whose 98% PrI intersects this zone requires manual review.
Notice how the positive and negative densities start overlapping starting with the second iteration of Session 2. Meanwhile the distribution of the records to be reviewed shrink and shift towards the negative side, as positive records get found and labeled.

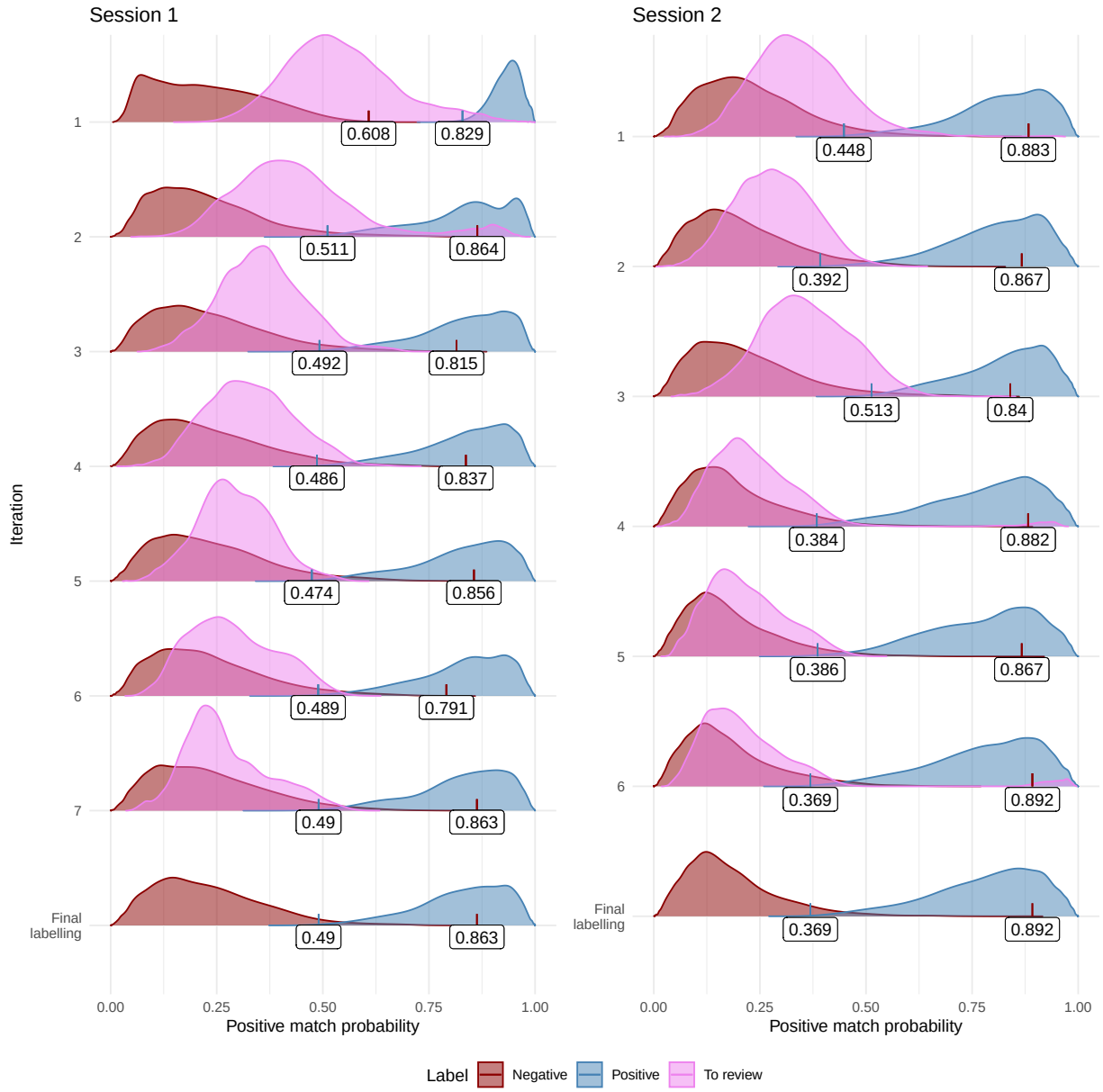

Figure 1. Mixture predictive distribution of the probability of a positive match, grouped by labelling status.

List of terms relevant for prediction

In table 1 and 2 are listed the 50 more relevant terms used by the BART algorithm to discriminate between positive and negative records, for Session 1 and 2. Term importance ("Value" in the tables) is defined as the proportion over 10 000 posterior trees in which a term was used divided by the standard deviation of this value among each model repetition. The symbol "|" indicate synonyms while terms joined by "&" represent terms that are co-present in a document component The component in which the term was used is reported in the leftmost column.
Next to it, we added the relative risk of a positive match (RR) and statistical relevancy (measured as standard errors, s.e.) of the terms estimated through a Poisson regression, to evaluate its linear correlation with the probability of a positive match. A strong BART score with a low regression score identify terms whose effect is highly non linear (e.g., they are relevant only in the presence of other terms).

Hyperparameters grid search

To select the engine hyperparameter set which would maximize sensitivity and efficiency, we set up a comprehensive grid search by running the algorithm on a fully labeled dataset of 1200 records. Using a partition tree algorithm we grouped the searches in a number of clusters with similar performance and selected the best parameter set in the best cluster.
Table 3 shows the best parameter set and their performance for each cluster while figure 2 displays the conditional impact on performance of each parameter.

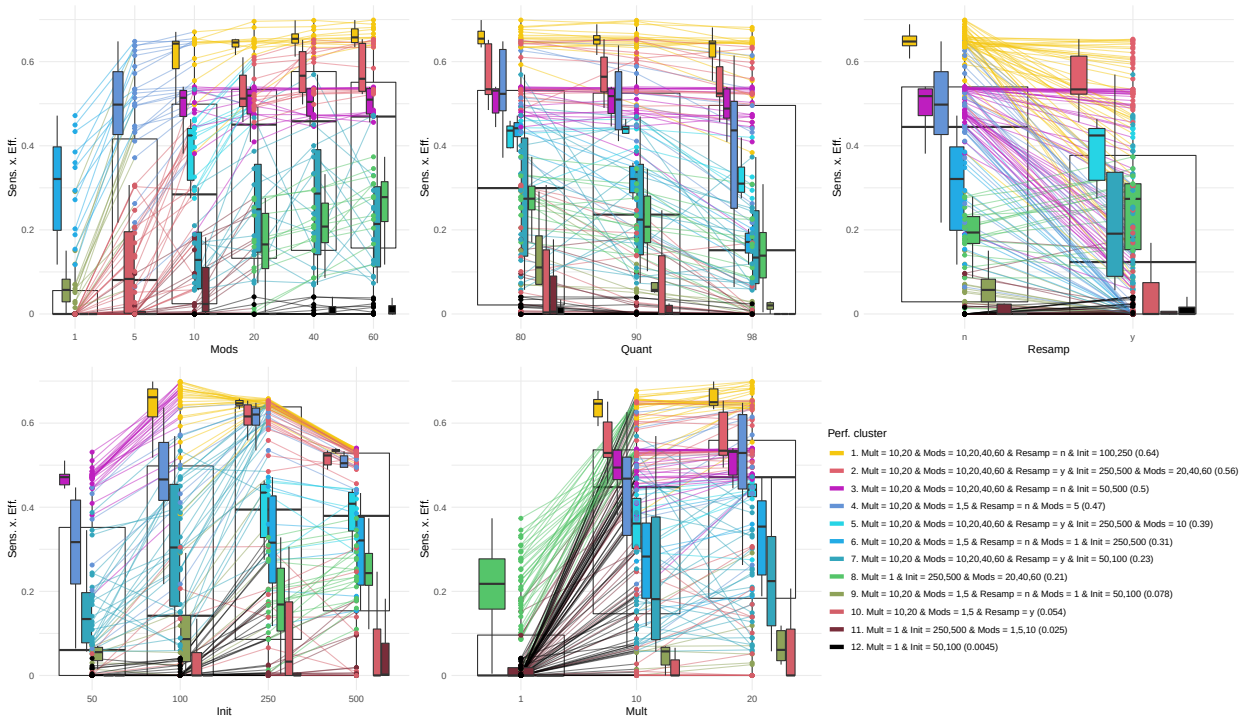

Figure 2. Performance clusters and impact of single hyperparameters on engine performance. The performance is measured as Sensitivity x Efficiency. Each cluster is color coded. Mods: num. of models in the ensemble; Quant: uncertainty zone interval to trigger manual review; Resamp: bootstrap resampling; Init: num. of initial training records; Multi: oversampling multiplier of positive matches.

Table 1. Term importance at the end of Session 1.

Component	Term	Value (on 10K trees)	RR	Statistic (s.e.)
Keyword	Patient Transport	61.2	99.1	21.3
Abstract	Transfer	57.0	22.5	15.4
Title	Network	56.5	18	14.2
Abstract	Network & Patient	54.2	26.3	15.2
Author	Donker T	53.5	159	16.5
Abstract	Worker	50.0	0.421	-1.21
Keyword	Hospitals	49.8	27.8	16.5
Abstract	Movement	47.8	27.2	15
Title	Spread	46.6	16.2	12.1
Abstract	Facility	45.0	19.6	14.8
Keyword	Orange County	44.3	199	17.2
Abstract	Conduct	42.6	0.221	-2.57
Abstract	Patient	42.0	27.6	7.23
Abstract	Perform	41.9	0.342	-2.55
Title	Hospital	39.0	12.5	12.5
Abstract	Regional	38.9	21.7	14.9
Abstract	Agent	38.1	4.36	6.28
Abstract	California	37.3	38	12.6
Title	Transfer	36.6	27	11.8
Keyword	Patient Transfer | Patient Transfer & Patient Transport	36.6	164	2
Abstract	Finding	33.2	0.372	-2.35
Title	Outbreak	32.9	3.4	3.51
Abstract	Collect	32.4	0.408	-1.95
Title	Regional	32.3	44.2	14.2
Abstract	Network	31.6	12.8	11.7
Abstract	Resistant	31.6	11	11.2
Abstract	Outcome	31.2	0.178	-2.95
Abstract	Discharge	31.1	9.99	9.02
Abstract	2014	30.3	0.588	-1.04
Abstract	Practice	29.5	0.508	-1.33
Abstract	Culture	28.9	0.378	-1.66
Abstract	Positive	28.8	0.346	-2.52
Abstract	Gene	28.3	2.4e-07	-0.0415
Abstract	Disease	28.0	0.365	-3.7
Keyword	Enterococci	27.5	25.2	6.33
Abstract	Month	27.3	0.288	-2.44
Abstract	Healthcare & Facility	26.2	34.9	17
Abstract	Prevalence	26.2	4.08	6.69
Abstract	Effort	26.1	6.64	8.57
Abstract	Length	25.3	6.1	6.45
Keyword	System	25.0	11.9	3.47
Abstract	Laboratory	24.4	0.442	-1.39
Keyword	Resistant Staphylococcus Aureus	24.4	22.8	11.2
Abstract	Clinical	23.6	0.421	-2.9
Abstract	Dataset	22.5	8.75	6.5
Abstract	Development	22.2	0.148	-2.68
Abstract	Hand	22.0	0.822	-0.335
Keyword	Pathogen Transmission	21.8	67.9	7.2
Keyword	Cross Infection & Humans & Transmission	21.7	31.1	15.5
Abstract	Flow	21.4	4.07	4.56

Table 2. Term importance at the end of Session 2.

Component	Term	Value (on 10K trees)	RR	Statistic (s.e.)
Keyword	Patient Transport	61.2	99.1	21.3
Abstract	Transfer	57.0	22.5	15.4
Title	Network	56.5	18	14.2
Abstract	Network & Patient	54.2	26.3	15.2
Author	Donker T	53.5	159	16.5
Abstract	Worker	50.0	0.421	-1.21
Keyword	Hospitals	49.8	27.8	16.5
Abstract	Movement	47.8	27.2	15
Title	Spread	46.6	16.2	12.1
Abstract	Facility	45.0	19.6	14.8
Keyword	Orange County	44.3	199	17.2
Abstract	Conduct	42.6	0.221	-2.57
Abstract	Patient	42.0	27.6	7.23
Abstract	Perform	41.9	0.342	-2.55
Title	Hospital	39.0	12.5	12.5
Abstract	Regional	38.9	21.7	14.9
Abstract	Agent	38.1	4.36	6.28
Abstract	California	37.3	38	12.6
Title	Transfer	36.6	27	11.8
Keyword	Patient Transfer | Patient Transfer & Patient Transport	36.6	164	2
Abstract	Finding	33.2	0.372	-2.35
Title	Outbreak	32.9	3.4	3.51
Abstract	Collect	32.4	0.408	-1.95
Title	Regional	32.3	44.2	14.2
Abstract	Network	31.6	12.8	11.7
Abstract	Resistant	31.6	11	11.2
Abstract	Outcome	31.2	0.178	-2.95
Abstract	Discharge	31.1	9.99	9.02
Abstract	2014	30.3	0.588	-1.04
Abstract	Practice	29.5	0.508	-1.33
Abstract	Culture	28.9	0.378	-1.66
Abstract	Positive	28.8	0.346	-2.52
Abstract	Gene	28.3	2.4e-07	-0.0415
Abstract	Disease	28.0	0.365	-3.7
Keyword	Enterococci	27.5	25.2	6.33
Abstract	Month	27.3	0.288	-2.44
Abstract	Healthcare & Facility	26.2	34.9	17
Abstract	Prevalence	26.2	4.08	6.69
Abstract	Effort	26.1	6.64	8.57
Abstract	Length	25.3	6.1	6.45
Keyword	System	25.0	11.9	3.47
Abstract	Laboratory	24.4	0.442	-1.39
Keyword	Resistant Staphylococcus Aureus	24.4	22.8	11.2
Abstract	Clinical	23.6	0.421	-2.9
Abstract	Dataset	22.5	8.75	6.5
Abstract	Development	22.2	0.148	-2.68
Abstract	Hand	22.0	0.822	-0.335
Keyword	Pathogen Transmission	21.8	67.9	7.2
Keyword	Cross Infection & Humans & Transmission	21.7	31.1	15.5
Abstract	Flow	21.4	4.07	4.56

Table 3. Hyperparameter clusters and best cluster subsets. For each cluster, the defining rules and mean Sens. x Eff. is shown, followed by the per-cluster best set results.

Cluster (mean score)	Num. iterations	Positive matches	Reviewed records	Sensitivity	Efficiency	Score (Sens. x Eff.)	Num. ensemble models	Uncertainty interval	Resampling	Num. initial training records	Positives oversampling multiplier
1. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = n & Init = 100,250 (0.64)	6	75 / 82	283 / 1200	91.5%	76.4%	0.699	60	80	n	100	20
2. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 250,500 & Mods = 20,40,60 (0.56)	6	76 / 82	354 / 1200	92.7%	70.5%	0.653	60	90	y	250	20
3. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = n & Init = 50,500 (0.5)	7	79 / 82	527 / 1200	96.3%	56.1%	0.54	10	80	n	500	20
4. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 5 (0.47)	7	74 / 82	338 / 1200	90.2%	71.8%	0.648	5	80	n	250	20
5. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 250,500 & Mods = 10 (0.39)	5	81 / 82	637 / 1200	98.8%	46.9%	0.463	10	90	y	250	20
6. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 1 & Init = 250,500 (0.31)	7	81 / 82	627 / 1200	98.8%	47.8%	0.472	1	80	n	250	20
7. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 50,100 (0.23)	6	79 / 82	491 / 1200	96.3%	59.1%	0.569	40	80	y	100	10
8. Mult = 1 & Init = 250,500 & Mods = 20,40,60 (0.21)	5	81 / 82	746 / 1200	98.8%	37.8%	0.374	60	80	y	500	1
9. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 1 & Init = 50,100 (0.078)	5	81 / 82	846 / 1200	98.8%	29.5%	0.291	1	80	n	100	20
10. Mult = 10,20 & Mods = 1,5 & Resamp = y (0.054)	5	82 / 82	832 / 1200	100%	30.7%	0.307	5	80	y	250	20
11. Mult = 1 & Init = 250,500 & Mods = 1,5,10 (0.025)	5	82 / 82	981 / 1200	100%	18.2%	0.182	10	90	y	500	1
12. Mult = 1 & Init = 50,100 (0.0045)	6	82 / 82	1151 / 1200	100%	4.08%	0.0408	20	80	y	50	1