Can ‘machine learning’ improve our understanding of non-response in Understanding Society and means of tacking it?

Stephen McKay

4 July 2019

Non-response (attrition)

- Model of non-response
    - Generally logistic regression, sometimes weighting class method (e.g. CHAID)
- Calculate probability of taking part in this wave, conditional on having been in the past wave
    - (non-monotone response patterns an issue)
- Generate a weight as inverse of the probability of taking part

Approach of different surveys

Understanding Society	GSOEP	BHPS	HILDA
.. Such as age, Gender Marital status Employment status Hh size Presence of children Hh spending on food Consideration of use of environmental energy, Among others	Household moved Large city Age (of head), female (head) Separation/divorce Change of interviewer; N-past-ints with interviewer Low income Item non-response on income Expect to lose job East Berlin	Weighting class method (CHAID). Age Sex Employment status Race Qualifications Organisational membership Region Tenure Cars Consumer durables	Age Sex Marital status “ability of speak English” Employment status Hours worked N kids Country of birth Highest education Health status Relationship in household Likelihood of moving Past moves

Models of non-response

- Logistic regression (or probit, ‘robit’, among others)
- Can be used to predict, but not generally tested for predictive power. Generally interested in the coefficients.
- Tend to be looking at within-sample, not going beyond (esp with stepwise methods); not looking for interactions

- Class of ‘machine learning’ models are designed for prediction, though (arguably) less clear to understand

Data preparation

- Understanding Society (today w1-w2)
- Combining xwavedat, indresp, hhresp, callrec datasets (about 1600 variables)
- Variables with <=20 unique values regarded as categorical, with 0/1 variables created (about 7500 variables)
- Dropped variables with no/low variance (-> 1900 variables)
- Sample divided into two equal groups (N=24,000 in each)
  - Training data to develop models
  - Validation (test) data to test models, on unseen data

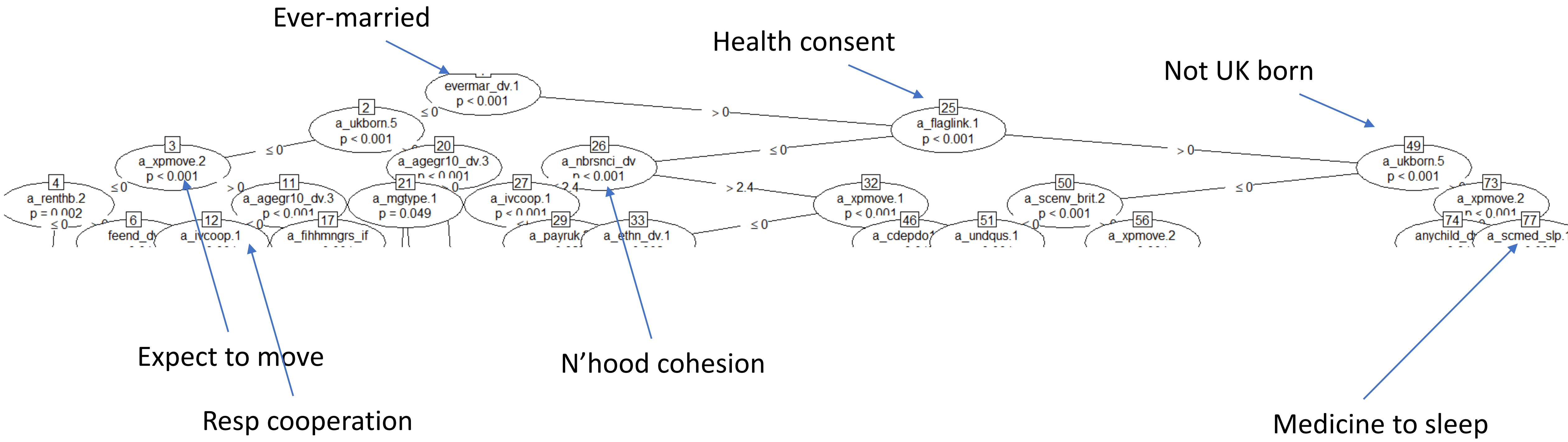
Logistic regression model to predict non-response (person level, w1-w2)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.372e-01	1.629e-01	3.297	0.000976	***
sex.2	1.377e-01	3.101e-02	4.439	9.05e-06	***
a_dvage	4.040e-02	5.211e-03	7.753	8.99e-15	***
a_dvagesq	-4.008e-04	5.059e-05	-7.922	2.33e-15	***
hhorig.7	-1.875e-01	5.158e-02	-3.635	0.000278	***
a_fihhmngrs1_dv	-7.507e-06	5.422e-06	-1.384	0.166208	
a_tenure_dv.1	9.148e-02	4.606e-02	1.986	0.047039	*
a_tenure_dv.3	-8.570e-02	5.558e-02	-1.542	0.123101	
a_tenure_dv.4	-5.839e-02	6.504e-02	-0.898	0.369369	
a_tenure_dv.6	-3.060e-01	5.934e-02	-5.157	2.51e-07	***
a_tenure_dv.7	-4.305e-01	6.476e-02	-6.647	3.00e-11	***
a_ncars.0	-6.694e-02	4.331e-02	-1.546	0.122163	
a_ncars.2	-2.628e-02	4.150e-02	-0.633	0.526639	
a_ncars.3	-2.286e-01	6.143e-02	-3.721	0.000199	***
a_gor_dv.2	-2.638e-02	6.614e-02	-0.399	0.689962	
a_gor_dv.3	-2.068e-01	6.909e-02	-2.994	0.002757	**
a_gor_dv.4	3.747e-03	7.336e-02	0.051	0.959266	
a_gor_dv.5	-2.501e-01	6.707e-02	-3.729	0.000192	***
a_gor_dv.6	-1.003e-01	6.938e-02	-1.445	0.148346	
a_gor_dv.7	-2.677e-01	6.399e-02	-4.184	2.86e-05	***
a_gor_dv.8	-1.165e-02	6.425e-02	-0.181	0.856103	
a_gor_dv.9	1.433e-01	7.607e-02	1.883	0.059634	.
a_gor_dv.11	-3.633e-01	7.169e-02	-5.068	4.03e-07	***
a_finnow.1	2.360e-01	7.567e-02	3.119	0.001817	**
a_finnow.2	1.379e-01	7.148e-02	1.929	0.053703	.
a_finnow.3	1.059e-01	7.050e-02	1.502	0.133009	
a_finnow.4	1.240e-01	7.935e-02	1.563	0.117996	
a_finfut.1	-3.857e-02	3.740e-02	-1.031	0.302422	
a_finfut.2	9.052e-03	4.449e-02	0.203	0.838792	
evermar_dv.2	-5.358e-01	4.280e-02	-12.519	< 2e-16	***
generation.1	-4.670e-01	4.823e-02	-9.684	< 2e-16	***
generation.2	-1.115e-01	5.418e-02	-2.057	0.039640	*
generation.3	1.182e-02	6.336e-02	0.187	0.852026	
a_pno.2	1.422e-02	3.591e-02	0.396	0.692008	
a_pno.3	4.723e-02	6.242e-02	0.757	0.449292	
a_sf1.1	-3.620e-02	4.492e-02	-0.806	0.420400	
a_sf1.3	3.759e-02	4.026e-02	0.934	0.350491	
a_sf1.4	-2.837e-02	5.054e-02	-0.561	0.574576	
a_sf1.5	-2.283e-01	6.538e-02	-3.493	0.000478	***

(single) Decision tree

- Older technique (e.g. ‘CHAID’), with more recent algorithms
- Seeks to divide the sample into groups that are like each other, and unlike other groups.
- Easy to understand (e.g. for policy-making?)
- Don’t tend to generalise well from training data to new data (‘overfitting’), unless ‘pruned’

Decision tree for w2 attrition



Random forest

- Multiple of trees (few hundred)
- Takes random groups of the sample to run
- Random sampling of groups of variables in each tree
- Large number of separate ‘trees’ generated
- Each tree may be quite weak on its own … but using all the trees makes for a strong overall model

Random Forest – most important variables

- Home ownership
- Housing Benefit receipt
- Married (Ever)
- Any children
- Agreed to health linkage
- Intend to move
- Ethnic boost sample
- GHQ score

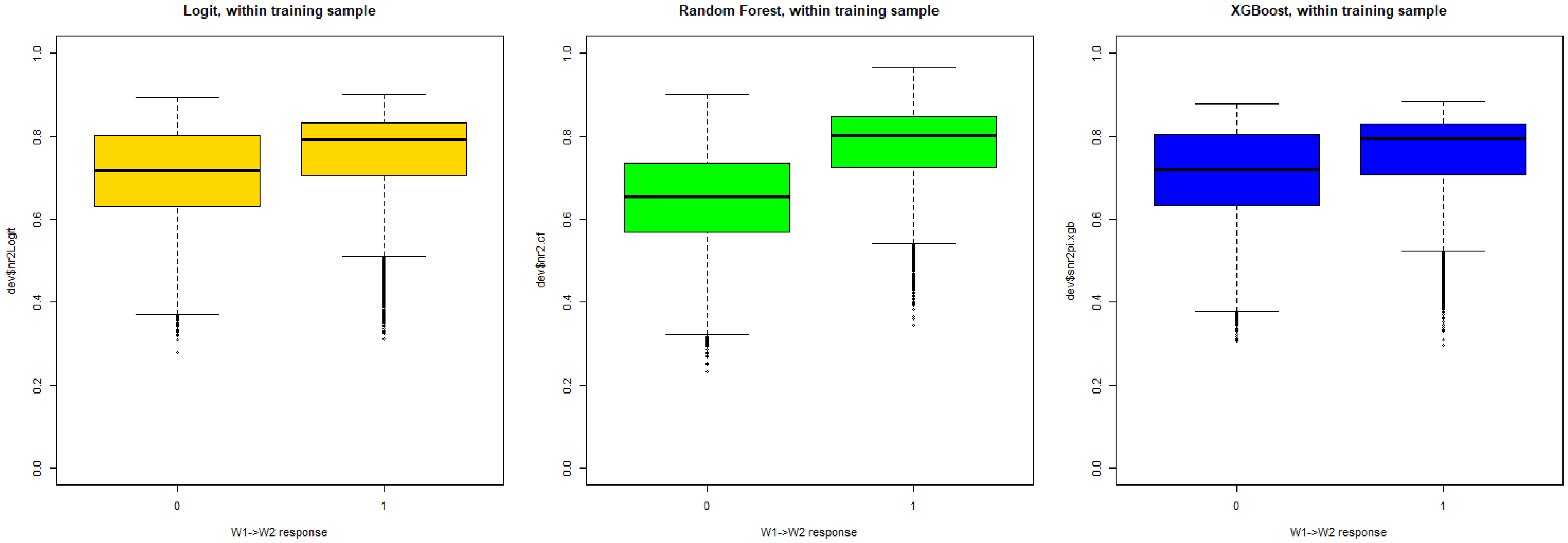
Boosting

- Rather than only independent trees, focus on the ‘hard’ cases
- Misclassified cases get higher weights, then put through same models
- Various approaches to how to ‘learn’
  - adaboost (Freund and Schapire)
  - Extreme gradient boosting (xgboost – 2015) – best single algorithm on competitive tests (?) e.g. Kaggle competitions

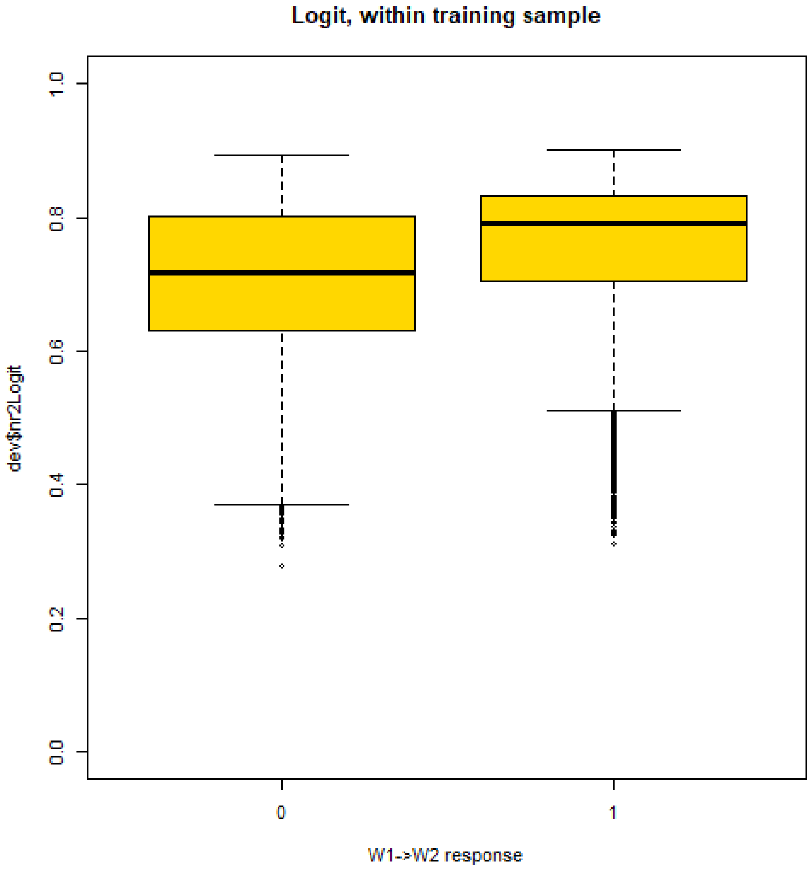
Xgboost – most important variables

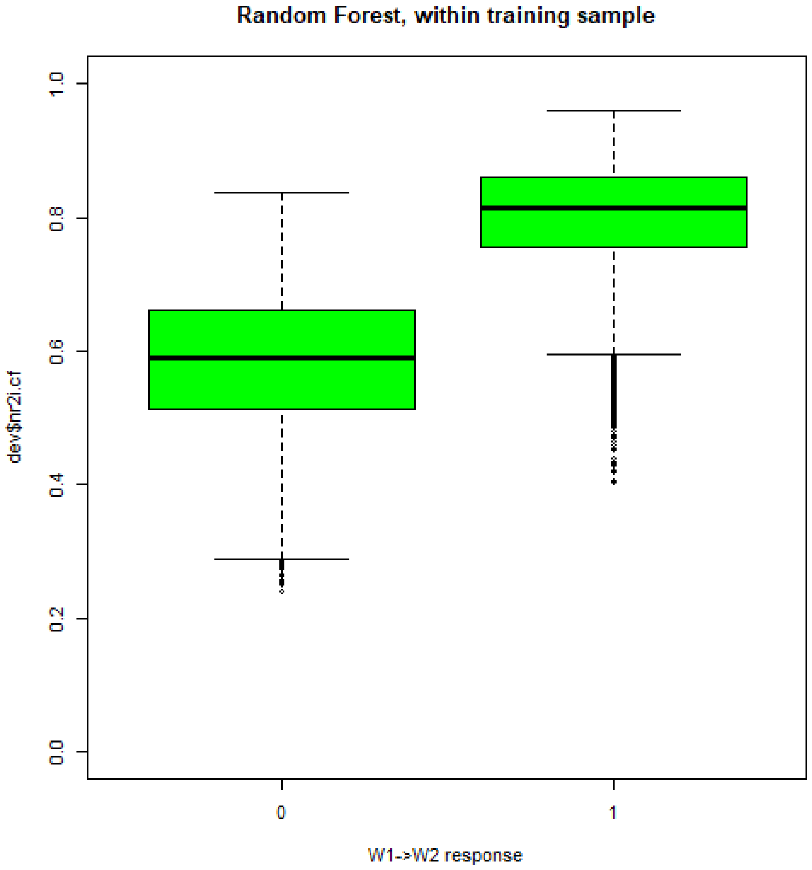
lmar1y_dv (first marriage year)	100.00	a_mvyr (year moved here)	26.57
ch1by_dv (first child birth year)	92.87	a_hhdenus_xw (x-section hh wgt)	22.28
a_xpmove.2 (expect to move nxt yr)	56.58	a_fihhmngrs_if (share imputed income)	22.15
a_indpxus_xw (x-section adult wgt)	56.39	paid.1 (father ethnic group)	21.52
a_flaglink.1 (health consent flag)	53.96	a_fiyrdia (interest income)	19.78
a_hscost (house purchase price)	53.82	a_istrtdathh (int start: hours)	18.87
a_hhresp_dv.3 (within-hh resp)	47.64	a_susp.1 (suspicious respondent)	18.83
a_nbrsnci_dv (N'hood cohesion)	35.51	a_plbornc (country of birth)	18.13
a_intnum (interviewer number)	34.57	a_hlphmwk..8 (parents help w homework)	17.62
a_scdoby4 (year of birth)	28.17		
a_ivcoop.1 (respondent co-operative)	26.60		

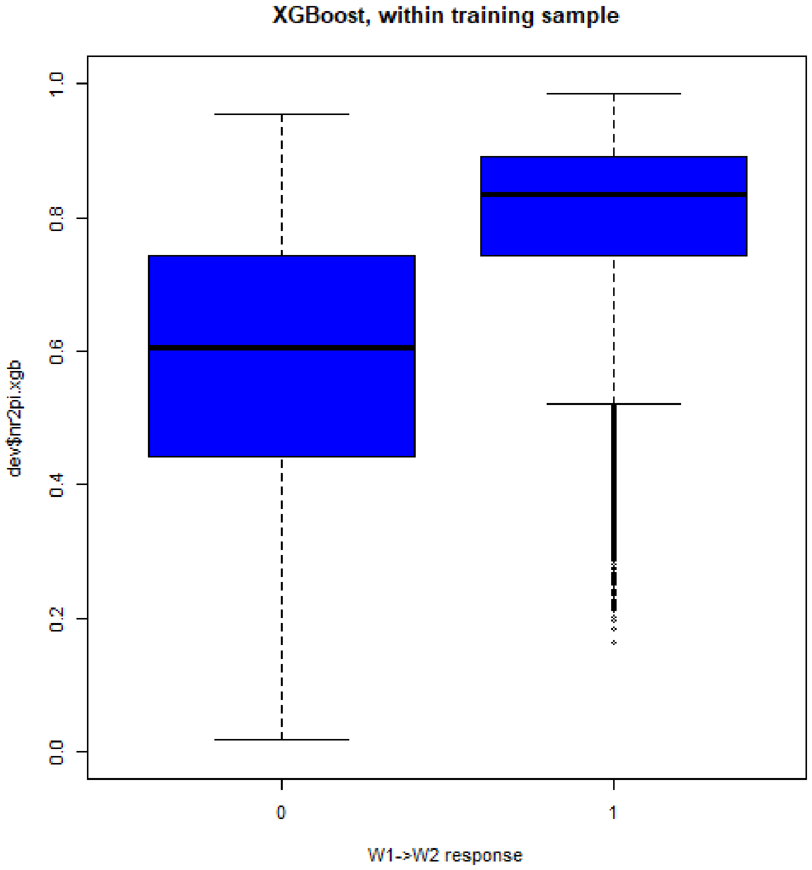
Predictions – same input variables



Predictions – ML with full dataset

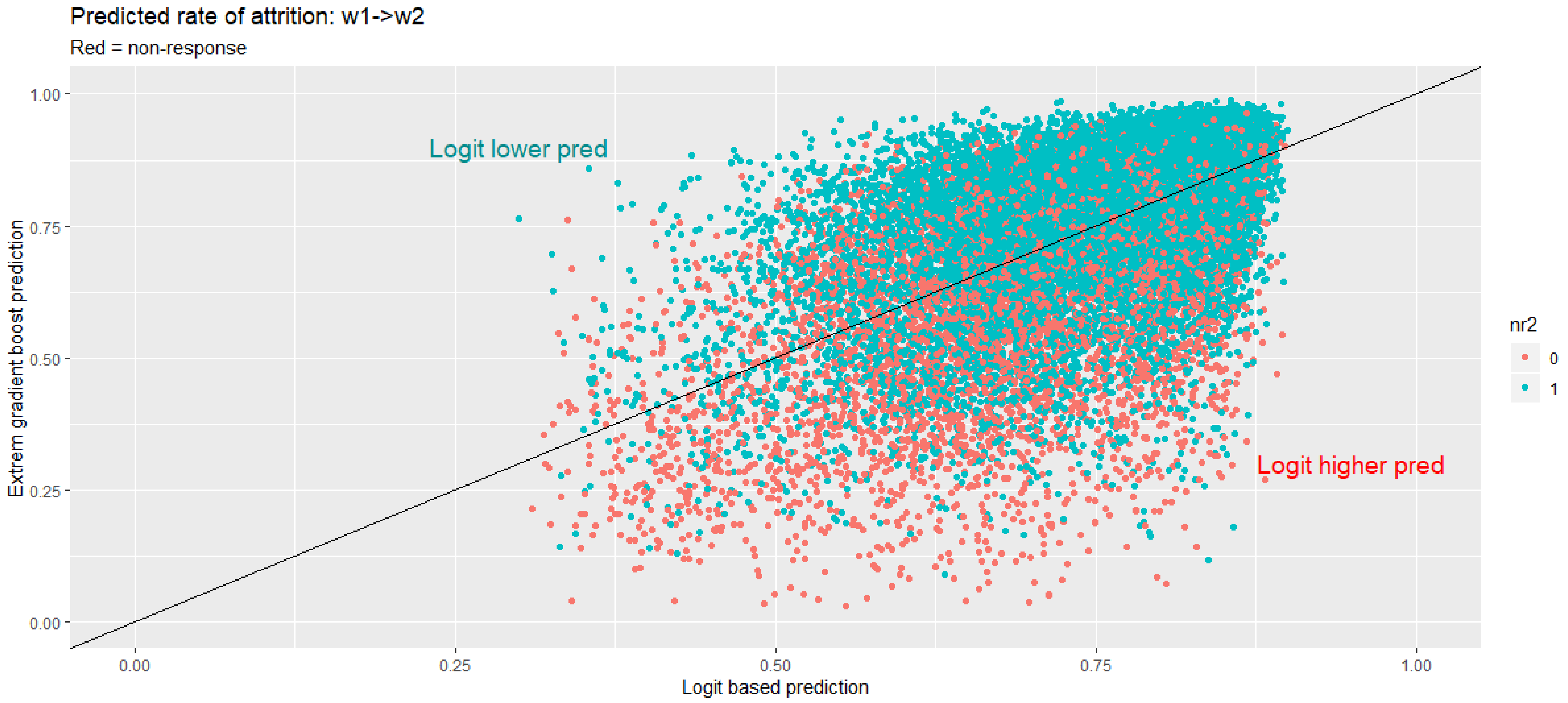






Predictive performance (mean sq error; correctly predicted)

Model	Training data	
Average (respondent)	0.190 (74%)	
Base logit (stepwise v similar, as is OLS)	0.179 (75%)	
Random Forest – same variables	0.151 (77%)	
XGBoost – same variables	0.178 (75%)	
Single decision tree – all variables	0.171 (76%)	
Random Forest – all variables	0.123 (80%)	
XGBoost – all variables	0.138 (80%)	

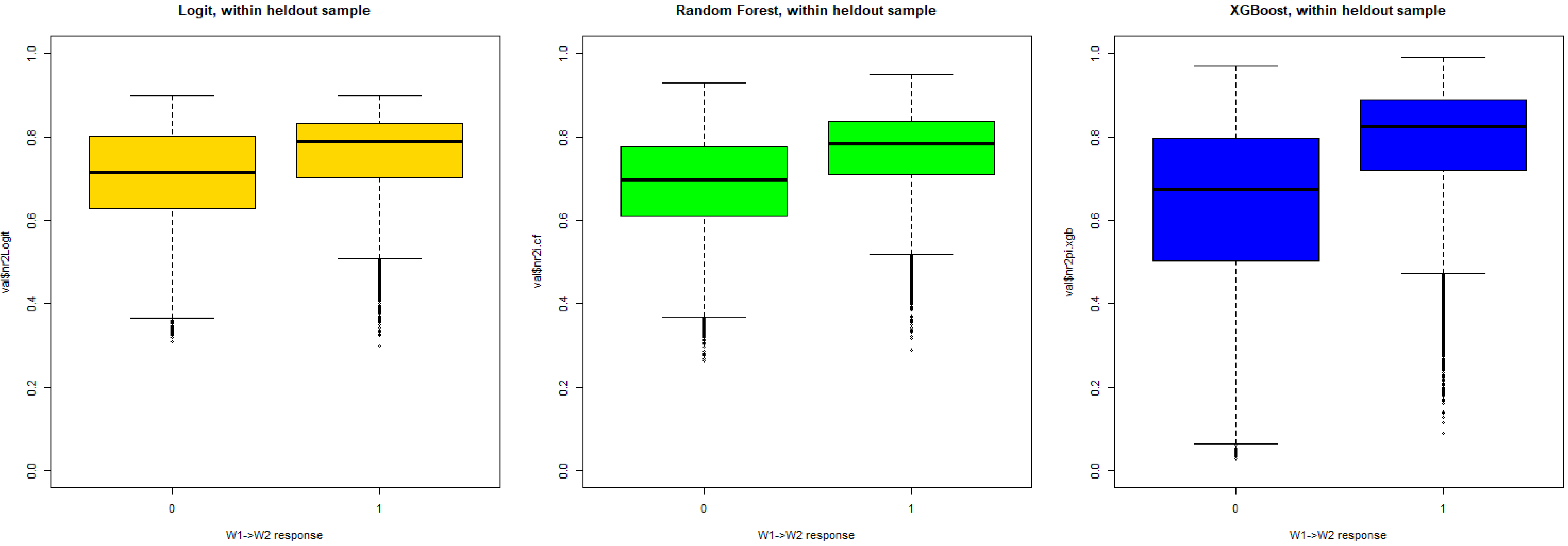


Predicted rate of attrition: w1->w2
Red = non-response
Logit lower pred
Logit higher pred
Extrem gradient boost prediction
Logit based prediction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
nr2
0
1

Predictive performance (mean sq error; correctly predicted)

Model	Training data	Held-out data
Average (respondent)	0.190 (74%)	0.190 (74%)
Base logit (stepwise v similar, as is OLS)	0.179 (75%)	0.181 (74%)
Random Forest – same variables	0.151 (77%)	0.180 (75%)
XGBoost – same variables	0.178 (75%)	0.180 (75%)
Single decision tree – all variables	0.171 (76%)	0.179 (75%)
Random Forest – all variables	0.123 (80%)	0.172 (75%)
XGBoost – all variables	0.138 (80%)	0.163 (77%)

Predictions – on held-out data



Effect of weights

	W1	W1 after attrition	Logit-based weights	RF-based weights	Xgb-based weights
Is female (%)	55.5	56.6	55.6	56.2	55.6
Mean (age)	46.0	47.3	46.0	46.5	46.1
SE (age)	0.12	0.13	0.14	0.14	0.14
(also considered: region, rural, income, health, depression, politics, ..)					
Range of predicted probabilities			0.279 – 0.900	0.24 – 0.960	0.018 – 0.985
Range –respondents			0.311 – 0.900	0.40 – 0.960	0.164 – 0.985
Kish design effect			1.026	1.013	1.060

Conclusions

- ML methods will more closely fit the probability of attrition … though not to any great extent on unseen data *except* best models with more data
- ML highlights some variables relating to ‘para data’ – process of taking part and permissions – that might be further explored for attrition models
- Distribution of weights from ML models *may* be more extreme (without truncating) than from standard statistical models
- Not (yet) found any advantages in terms of descriptive statistics with weights from ML models over existing practice