Table 1: Best Balanced Accuracy for Each Statistical Algorithm across 10 Folds during Model Selection

algorithm	feature set	feature type	resample	balanced accuracy
random_forest	passive	perc_raw	down_1	0.61
glmnet	active	raw	up_1	0.60
knn	passive	raw	down_1	0.58

Table 2: Classification Performance Metrics for Best Model across 100 Folds

metric	estimate
bal_accuracy	0.60
sens	0.46
spec	0.73
roc_auc	0.64
accuracy	0.72
ppv	0.09
npv	0.96

Results

Participant Characteristics

Model Selection

We optimized each statistical algorithm by tuning hyperparameter values and fitting models across several feature set combinations (i.e., active or passive, and type of feature engineering). Each model configuration was fit using a grouped 1x10 resampling method. Table 1 shows the best performing model (i.e., highest balanced accuracy) for each statistical algorithm. Figure 1 shows the model's performance in each held out fold.

Our top performing model, with the highest balanced accuracy, was a random forest statistical algorithm using passive features. To reduce the effects of optimization bias on our model evaluation of predictive performance, we refit the top performing model 100 times (grouped 10x10 resampling). We then averaged across performance estimates to get an estimate with low variance. This method gave us a balanced accuracy estimate of .60. Table 2 characterizes this model over several metrics appropriate for classification. Table 3 shows a confusion matrix where we can see how well the model predicts with negative cases (i.e., no lapse) compared to positive cases (i.e., lapses).

Since we did not have an independent held out test set we were not able to completely remove optimization bias. So, we performed a model comparison to assess our model's performance compared to a null model with no signal. A Bayesian correlated t-test revealed a posterior probability that the balanced accuracy of our model was above the Region of Practical Equivalence (ROPE) is .999 (Figure 2). This suggests there is a meaningful difference between our model and a null model.

In the appendix we show that our best performing model has variation in predictions for each individual participant. Figure A1 contains predictions that are predicted probabilities of a lapse. Each observation was held out 10 times and the figure shows the averaged probability across these 10 predictions. Actual lapses, are depicted in red.

Table 1

Table 2

Table 3

```
##            Truth
## Prediction     no    yes
##        no  143196   5386
##        yes  52314   4904
```

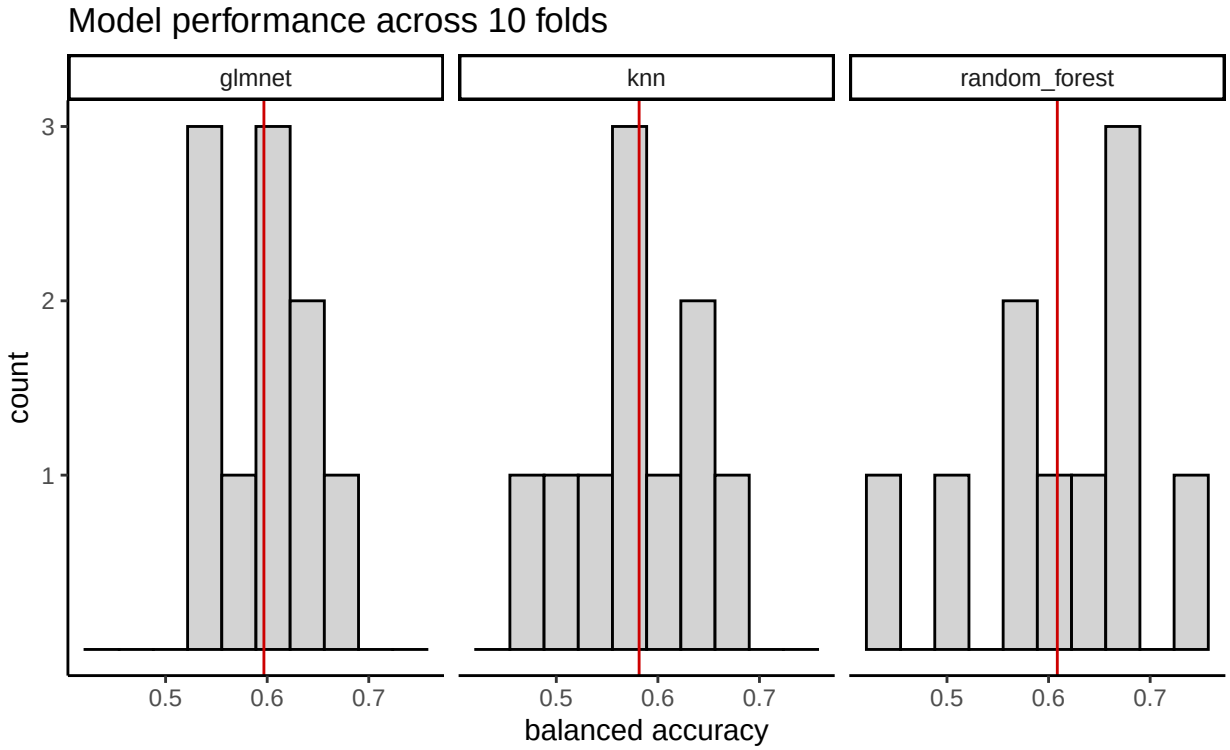


Figure 1. Model performance in each held out fold during model selection.

metric	estimate
bal_accuracy	0.60
sens	0.46
spec	0.73
roc_auc	0.64
accuracy	0.72
ppv	0.09
npv	0.96

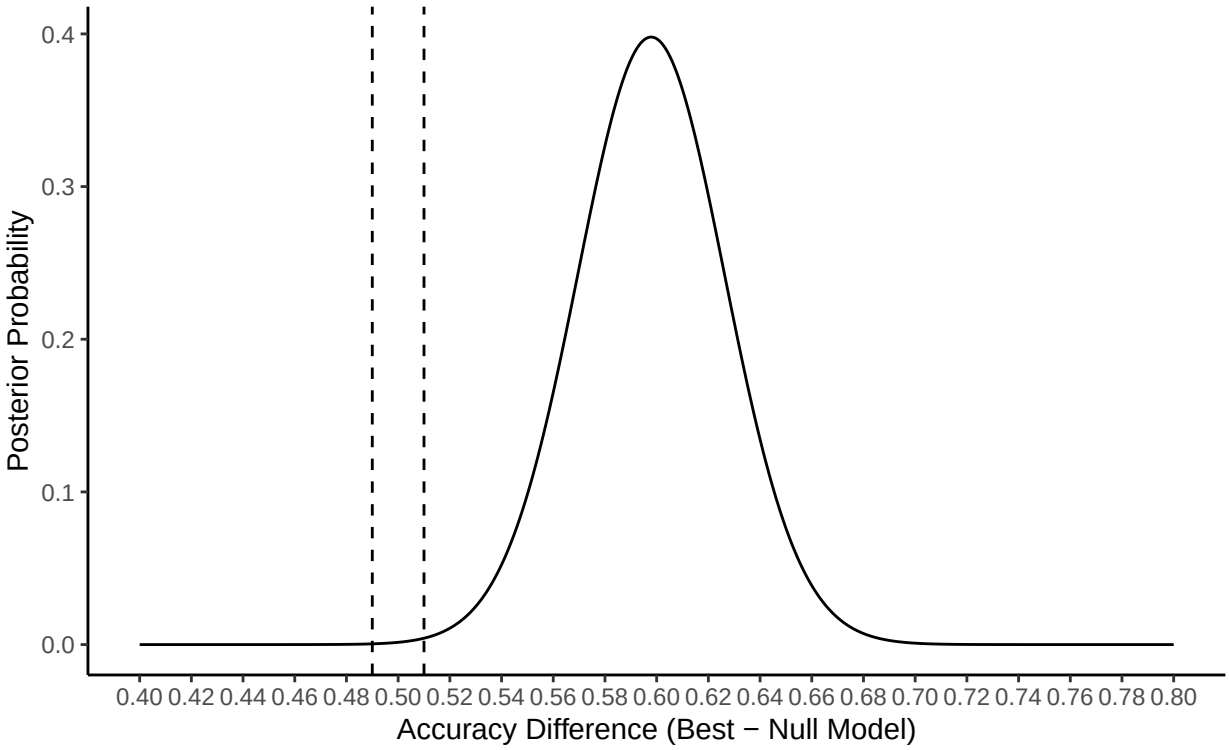


Figure 2. Model comparison of best model and null model.

Context Comparison

Our best performing active model using all the features available was a glmnet algorithm with a balanced accuracy of XX. Table 4 characterizes our best active model using various performance metrics. Figure 3 shows the balanced accuracy estimates for passive and active models over 100 fits.

A Bayesian correlated t-test revealed there were no differences between the best active and best passive models, ROPE = XX (Figure 4). This suggests we can predict lapses just as well with only passive features.

Table 4

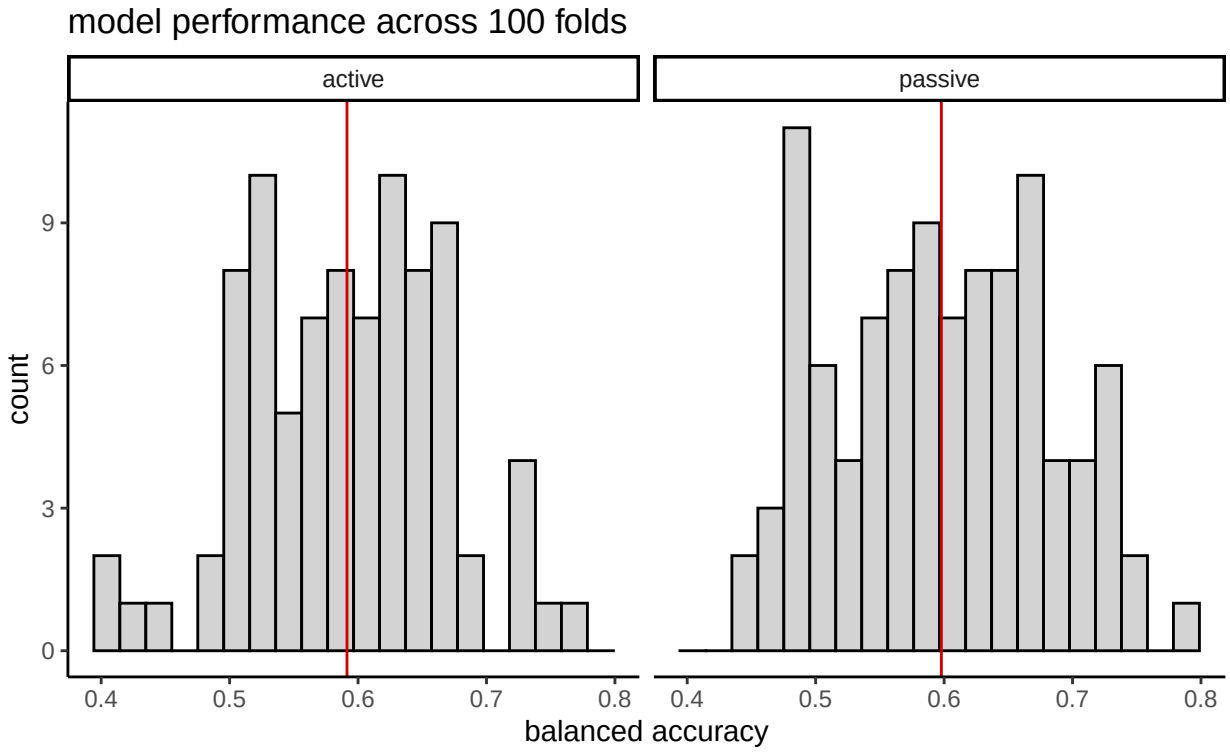


Figure 3. Histogram of model fits across 100 folds. (FIX: 14 folds still running so only 86 splits in active)

```
## Warning in cv_full - cv_compact: longer object length is not a multiple of
## shorter object length
```

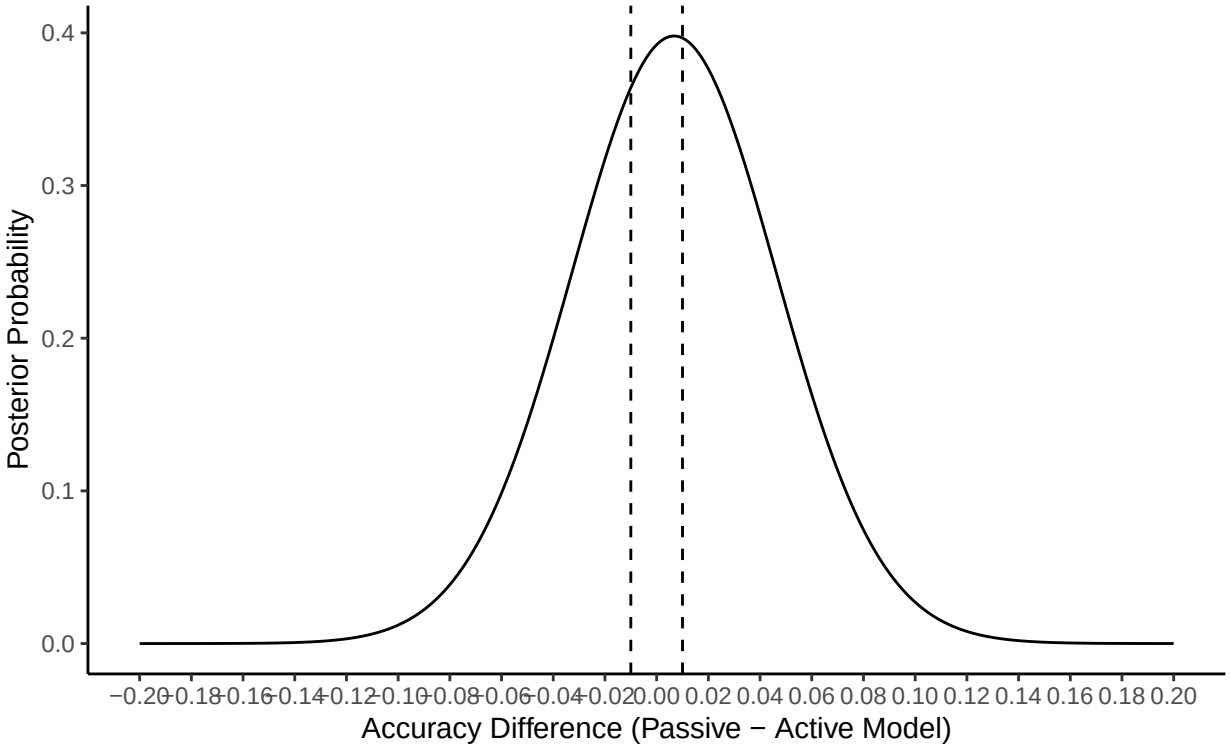


Figure 4. Model comparison of passive and active model.

Top Predictors

We conducted a permutation test of feature importance to see which features contributed most to our top performing model. Figure 4 shows which features when removed from the model had the greatest effect on performance. Figure 5 shows the top features for the best active model.

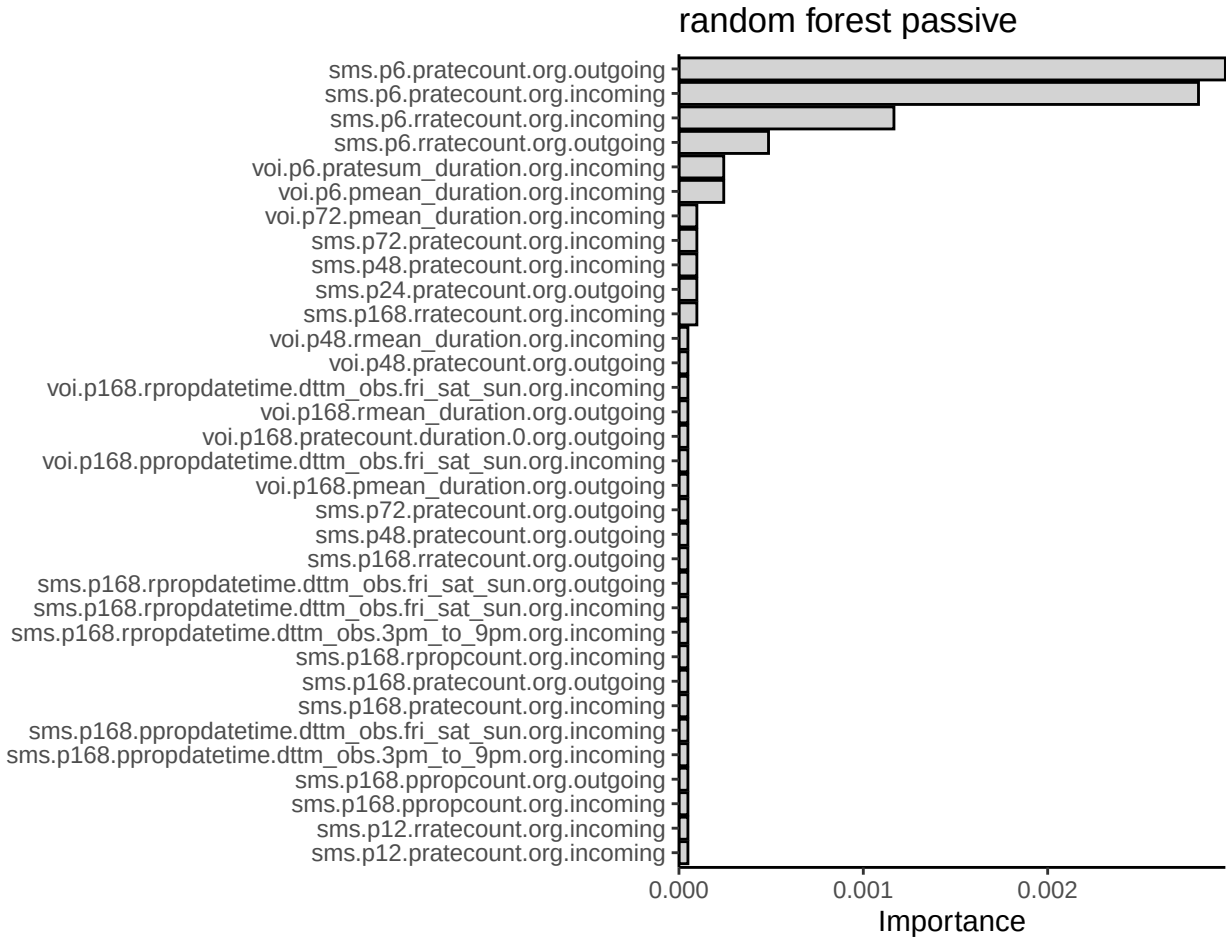
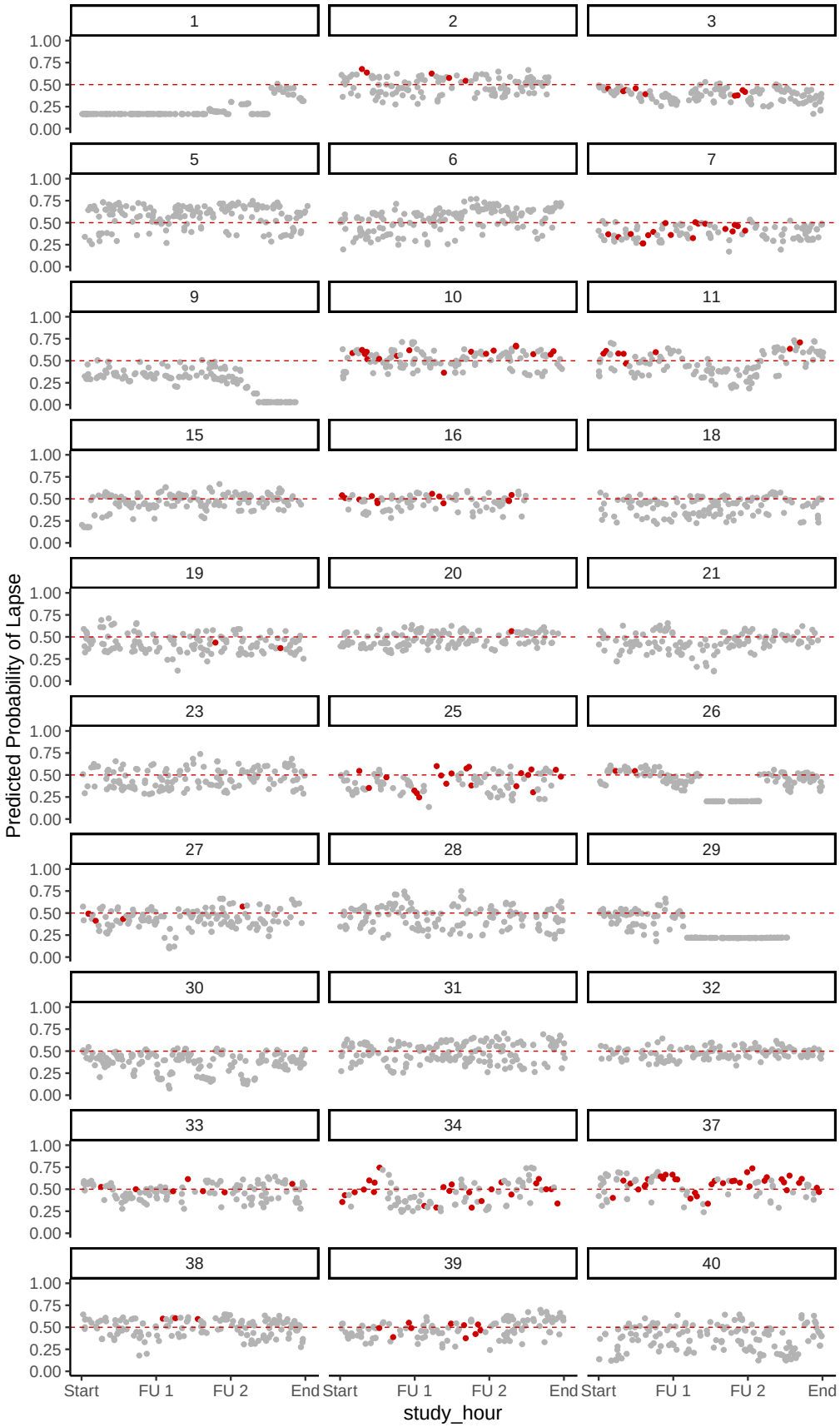
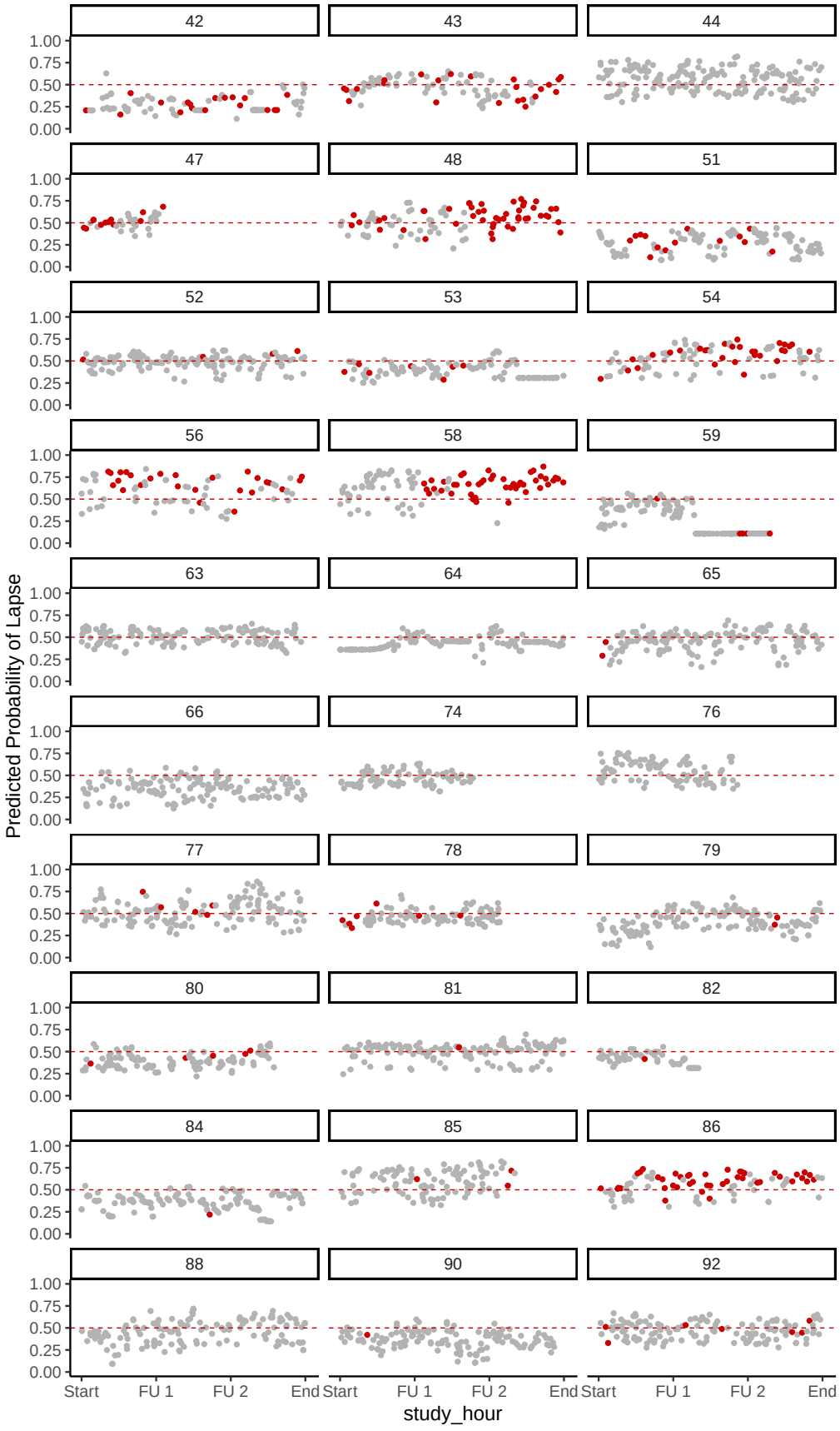


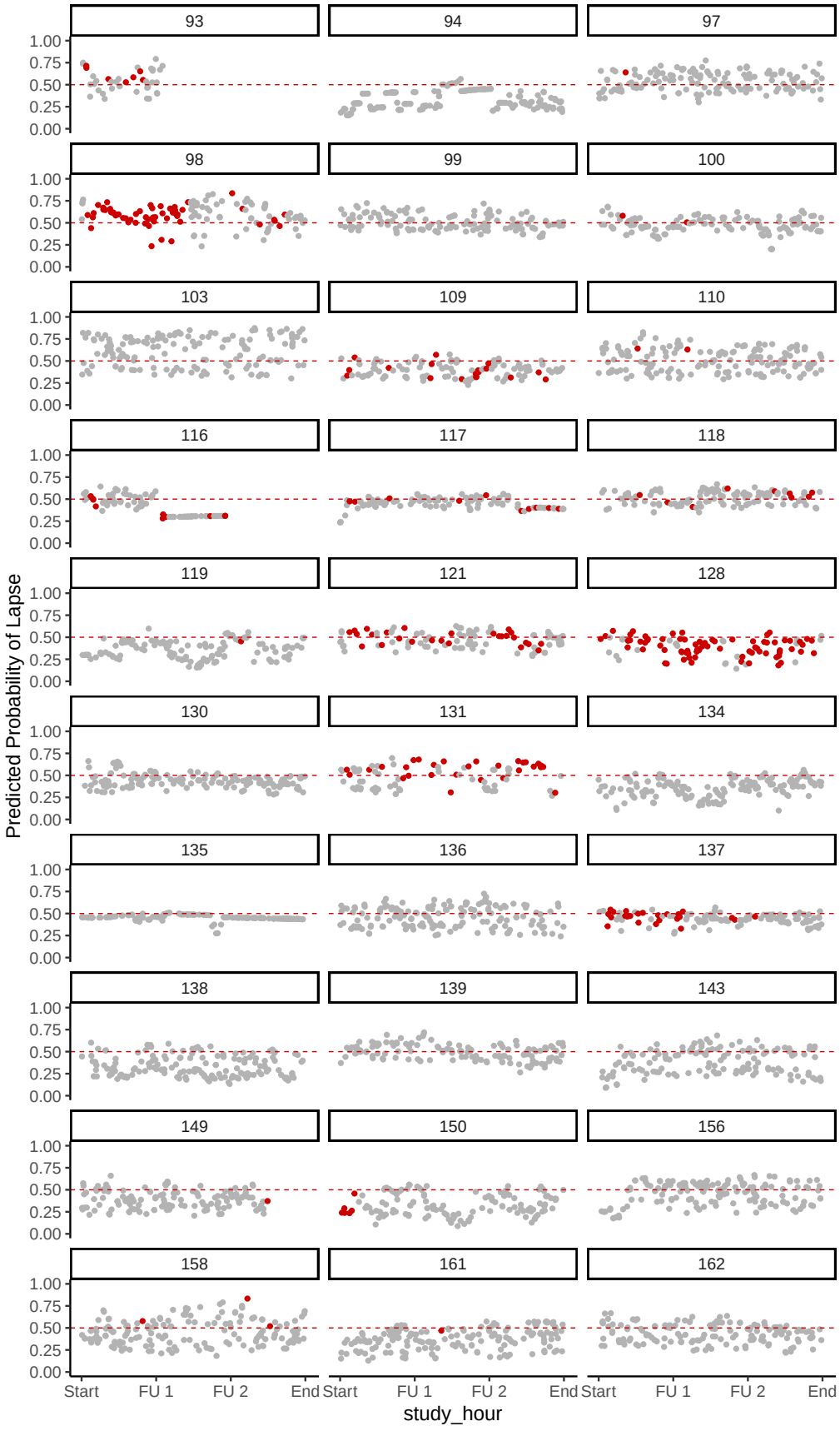
Figure 4. Feature importance scores for best performing model (Passive Random Forest).

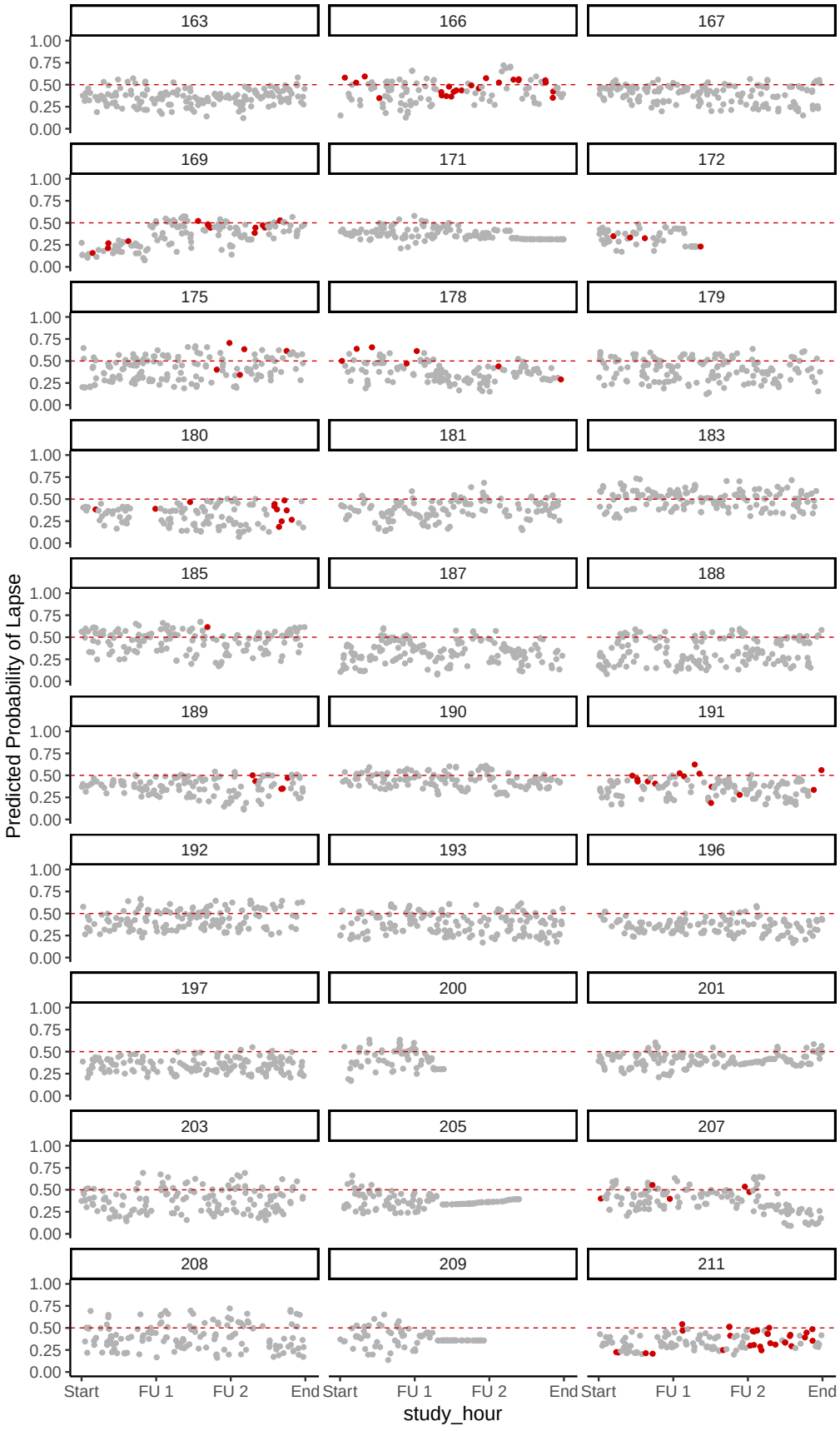
Figure 5. Feature importance scores for best performing active model (glmnet).

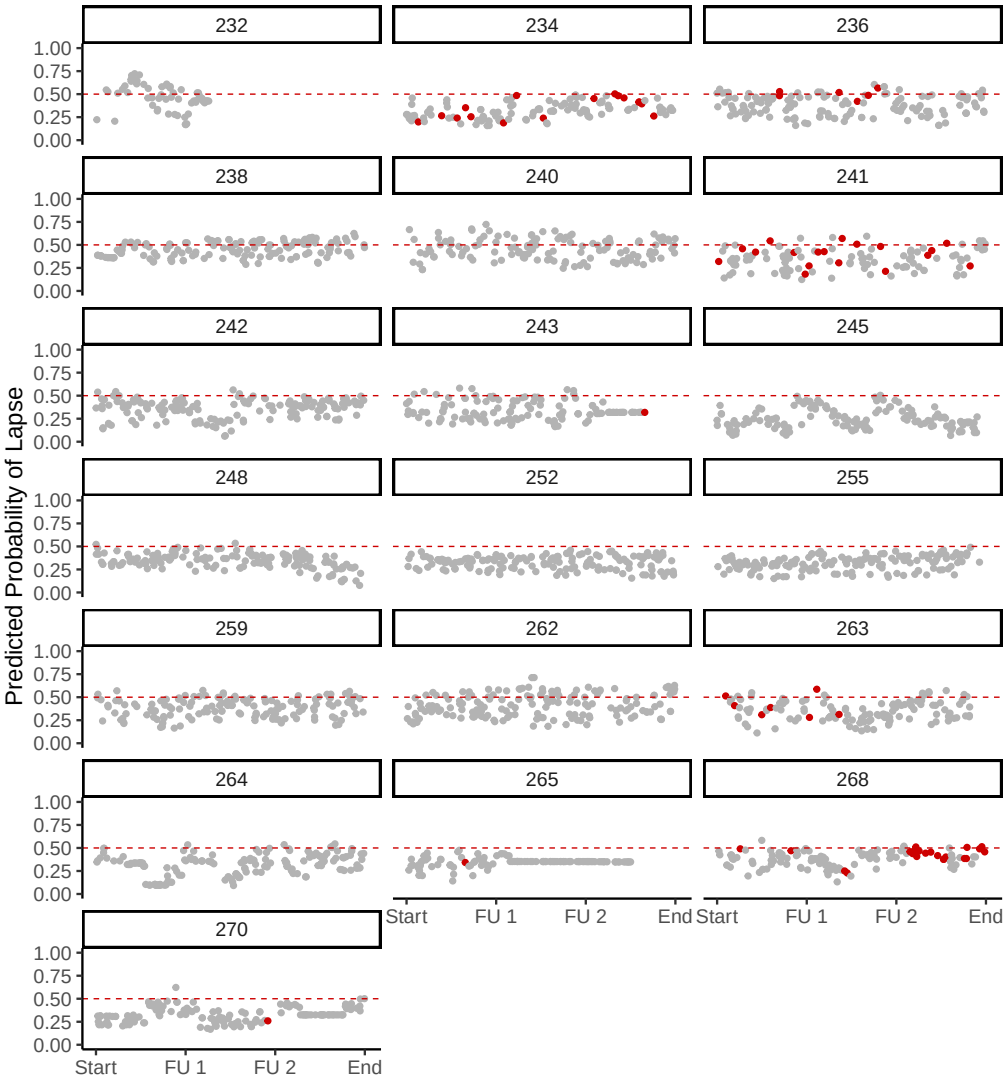
Appendix


1
2
3
5
6
7
9
10
11
15
16
18
19
20
21
23
25
26
27
28
29
30
31
32
33
34
37
38
39
40
1.00
0.75
0.50
0.25
0.00
Predicted Probability of Lapse
Start
FU 1
FU 2
End
study_hour


42
43
44
47
48
51
52
53
54
56
58
59
63
64
65
66
74
76
77
78
79
80
81
82
84
85
86
88
90
92
1.00
0.75
0.50
0.25
0.00
Predicted Probability of Lapse
Start
FU 1
FU 2
End
study_hour


93
94
97
98
99
100
103
109
110
116
117
118
119
121
128
130
131
134
135
136
137
138
139
143
149
150
156
158
161
162
Predicted Probability of Lapse
1.00
0.75
0.50
0.25
0.00
Start
FU 1
FU 2
End
study_hour


163
166
167
169
171
172
175
178
179
180
181
183
185
187
188
189
190
191
192
193
196
197
200
201
203
205
207
208
209
211
1.00
0.75
0.50
0.25
0.00
Predicted Probability of Lapse
Start
FU 1
FU 2
End
study_hour


232
234
236
238
240
241
242
243
245
248
252
255
259
262
263
264
265
268
270
Predicted Probability of Lapse
1.00
0.75
0.50
0.25
0.00
Start
FU 1
FU 2
End

study_hour

Figure A1. Predicted probabilities of lapse for each participant. A grouped 10x10 resampling method was used to obtain these probabilities. Known lapses are in red. The red dashed line represents the threshold for classifying a probability as a lapse (i.e., everything above the line was predicted to be a lapse).