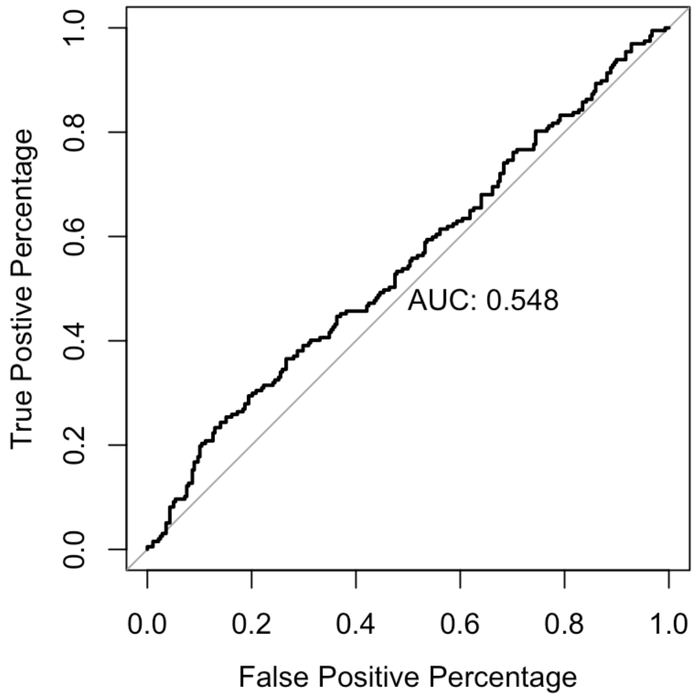
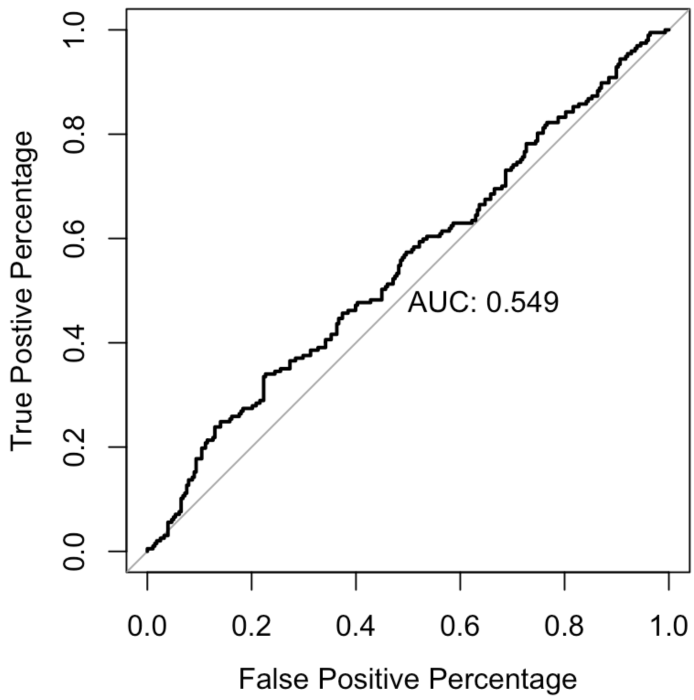
1a **Tao Xiaoying**

Model	1	2	3	4	5	6	7
AIC	2555	2556.3	2555.6	2574.7	2559.3	2562.3	2557.7
AUC	0.548	0.549	0.548	0.574	0.546	0.570	0.541

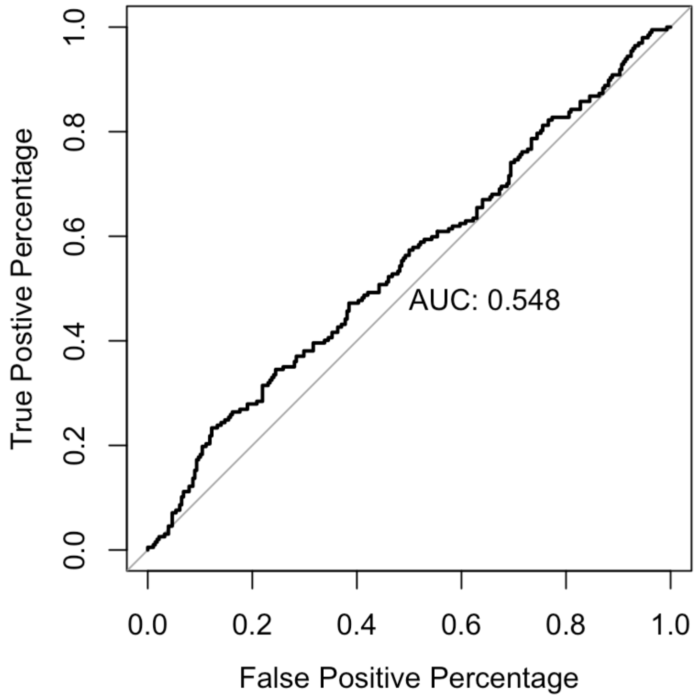
ROC of Model 1



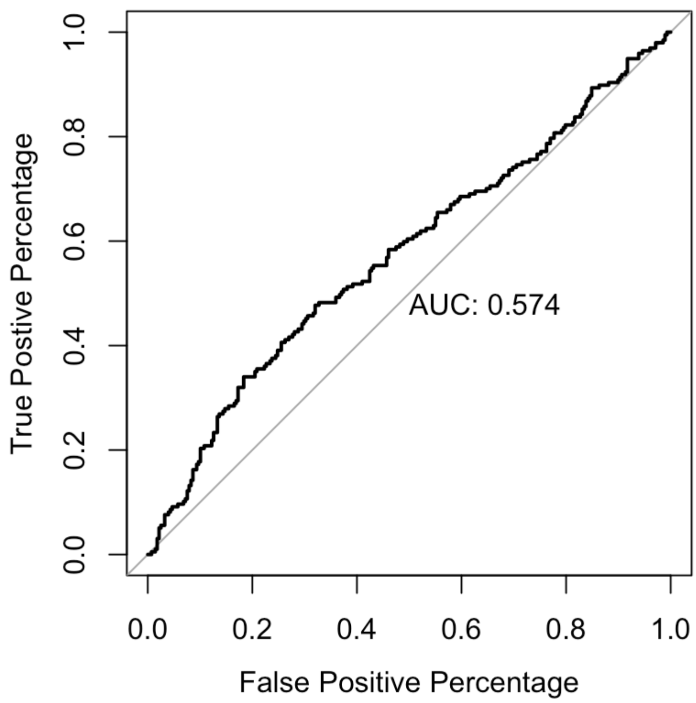
ROC of Model 2



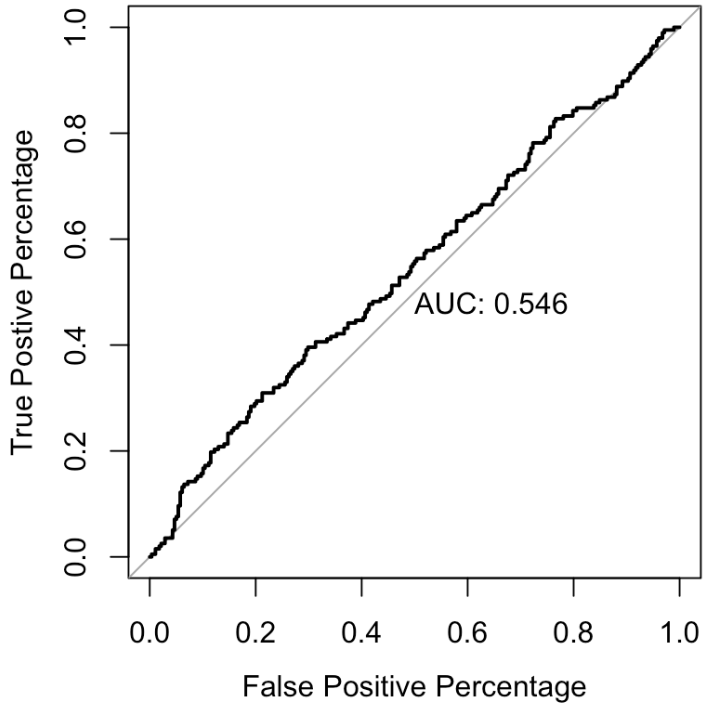
ROC of Model 3



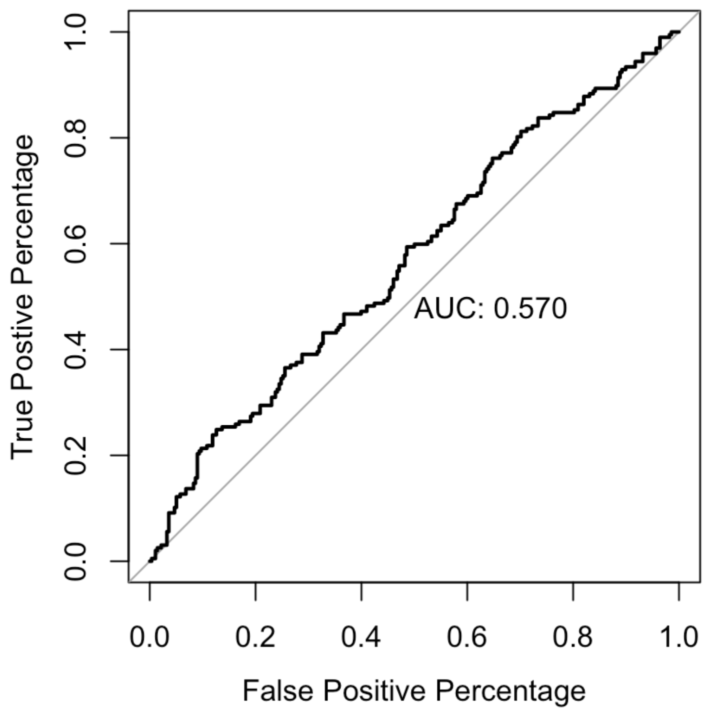
ROC of Model 4



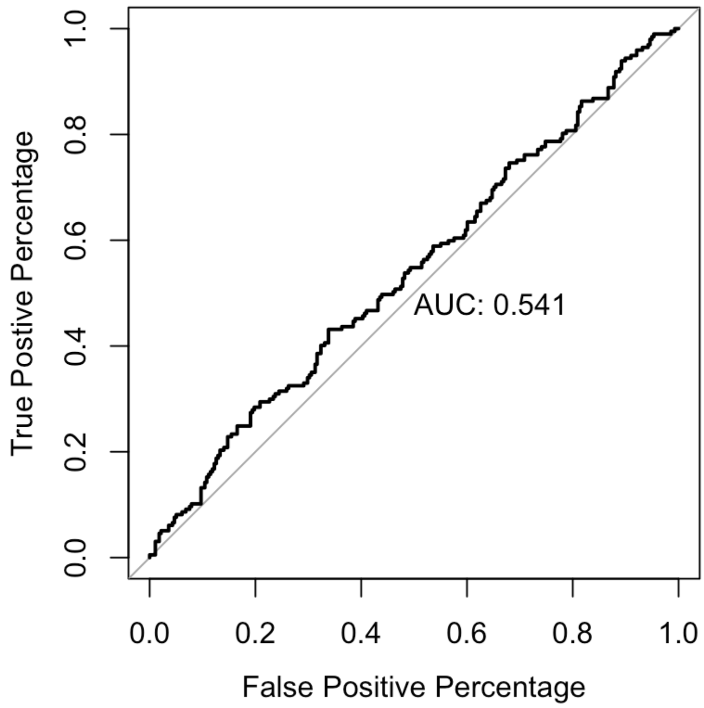
ROC of Model 5



ROC of Model 6



ROC of Model 7



Explanation:

According to information criteria (here it's AIC), model 1 is the best. The reason for this is that model 1 contains fewer variables.

By drawing ROC curve and calculating AUC, it's clear that model 4 performs the best. However, as there's no loss function, we are not sure if it's practically performing well.

1b

Confusion Matrix and Statistics of Forest 1

```
          Reference
Prediction   0   1
         0 225 141
         1  53  56

               Accuracy : 0.5916
                 95% CI : (0.5459, 0.6361)
    No Information Rate : 0.5853
    P-Value [Acc > NIR] : 0.4089

                  Kappa : 0.1001
```

Confusion Matrix and Statistics of Forest 2

```
          Reference
Prediction   0   1
         0 219 140
         1  59  57

               Accuracy : 0.5811
                 95% CI : (0.5352, 0.6258)
    No Information Rate : 0.5853
    P-Value [Acc > NIR] : 0.593

                  Kappa : 0.082
```

Confusion Matrix and Statistics of Forest 3

```
          Reference
Prediction   0   1
         0 207 129
         1  71  68

               Accuracy : 0.5789
                 95% CI : (0.5331, 0.6238)
    No Information Rate : 0.5853
    P-Value [Acc > NIR] : 0.6286

                  Kappa : 0.0938
```

Confusion Matrix and Statistics of Forest 4

```
          Reference
Prediction   0   1
         0 213 136
         1  65  61

               Accuracy : 0.5768
                 95% CI : (0.531, 0.6217)
    No Information Rate : 0.5853
    P-Value [Acc > NIR] : 0.6632

                  Kappa : 0.08
```

Explanation:

The first forest is the best at performance regarding the accuracy and other statistics, and the reason behind this can be that it has more variables included, thus having larger number of trees and larger number of splits.

2a

CATE

conditional independence: $(Y_i^1, Y_i^0) \perp\!\!\!\perp W_i | X_i$

the conditional expectation of the outcome ($\mu_w(X)$):

$$E[Y_i | X_i = x, W_i = w] \text{ with } W \in \{0,1\}$$

by taking the difference, lead to CATE:

$$\begin{aligned}\tau(X) &= \mu_1(X) - \mu_0(X) \\ &= E[Y_i | W_i = 1, X_i = x] - E[Y_i | W_i = 0, X_i = X] \\ &= E[Y_i^1 | W_i = 1, X_i = x] - E[Y_i^0 | W_i = 0, X_i = x] \\ &= E[Y_i^1 | X_i = x] - E[Y_i^0 | X_i = x] \\ &= E[Y_i^1 - Y_i^0 | X_i = x]\end{aligned}$$

ATE

Because it's randomized treatment, we assume $W_i \perp (Y_i^1, Y_i^0)$

In this case ATE = ATOT, because $E[Y_i | W_i = 1] = E[Y_i^1 | W_i = 1]$ $= E[Y_i^1]$, and $E[Y_i | W_i = 0] = E[Y_i^0]$.

$$\begin{aligned}ATE &= E[Y_i^1 - Y_i^0 \mid X_i = 1] \\ &= E[Y_i^1 | X_i = 1] - E[Y_i^0 | X_i = 1] \\ &= \sum_{i=1}^{N} (Y_i^1 | X_i = 1) - \sum_{i=1}^{N} (Y_i^0 | X_i = 1)\end{aligned}$$

2b
ATE

prediction: $\hat{\mu}(x)$

leaf: $L(x)$

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i : X_i \in L(x)\}} Y_i$$

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1,\ X_i \in L\}|} \sum_{\{i : W_i = 1, X_i \in L\}} Y_i - \frac{1}{|\{i : W_i = 0,\ X_i \in L\}|} \sum_{\{i : W_i = 0, X_i \in L\}} Y_i$$

$$ATE = \hat{\tau}(x) \cdot \frac{1}{10000}$$

CATE

$$CATE = \frac{\sum_{i=1}^{10000} m(X_i)(1 - m(X_i))\, E[Y_i^1 - Y_i^0 \mid X_i = x]}{\sum_{i=1}^{10000} m(X_i)(1 - m(X_i))}$$

where $m(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$

CATE on X_1

	$X_1 \geq 0$	$X_1 < 0$
τ	1.5	-1.5

CATE on X_1, X_2

	$X_1 \geq 0$	$X_1 < 0$
$X_2 \geq 0$	3	-1
$X_2 < 0$	1	-2