Usually, I use the following to generate some (here 12) covariates that could be correlated

```
library(FactoMineR)
n=279
library(clusterGeneration)
library(mnormt)
k=12
S=genPositiveDefMat("unifcorrmat",dim=k)
X=round(rmnorm(n,varcov=S$Sigma)+8,2)
rownames(X)=1:n
colnames(X)=LETTERS[1:k]
```

Then I need to generate some data, based on some covariates (5 out of 12), with various strengths

```
idx = sample(1:k,size=5)
u = sample(c(-(4:1),1:4),5)
beta = rep(0,k)
beta[idx] = u
U = X%*%beta
U = U-min(U)
U = U/max(U)*6-3
p = exp(( U))/(1+exp((U )))
Y = rbinom(n,size=1,prob=p)
df = data.frame(Y=as.factor(Y),X)
levels(df$Y)=levels=c("blue","red")
```

We can run a classification tree

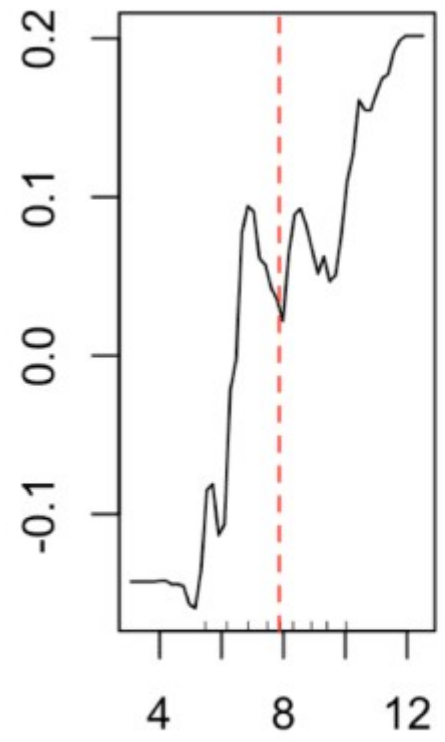
```
library(rpart)
arbre = rpart(Y~., data=df)
```

and a random forest,

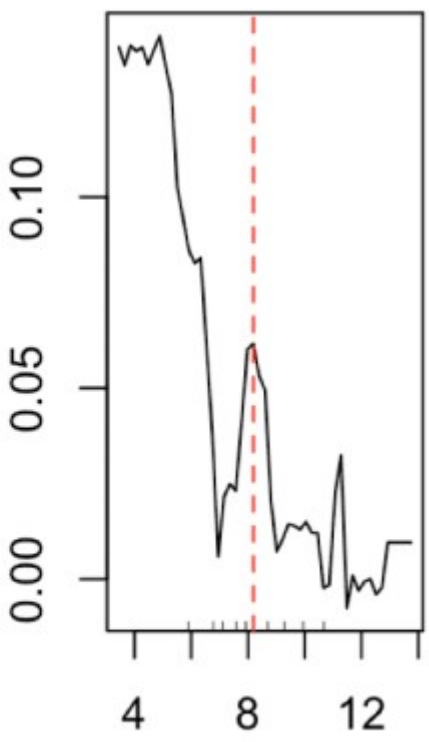
```
library(randomForest)
set.seed(1)
arbres = randomForest(Y~., data=df)
```

Here are the partial plots for 4 of the explanatory variables that actually have an impact

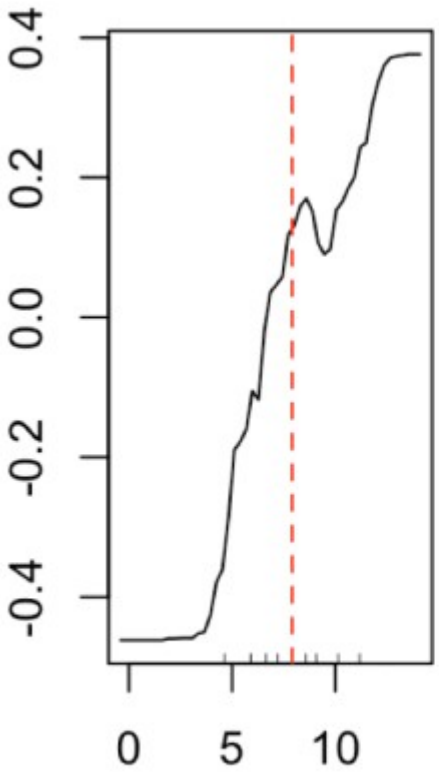
```
partialPlot(arbres,pred.data = df, x.var = "A")
```



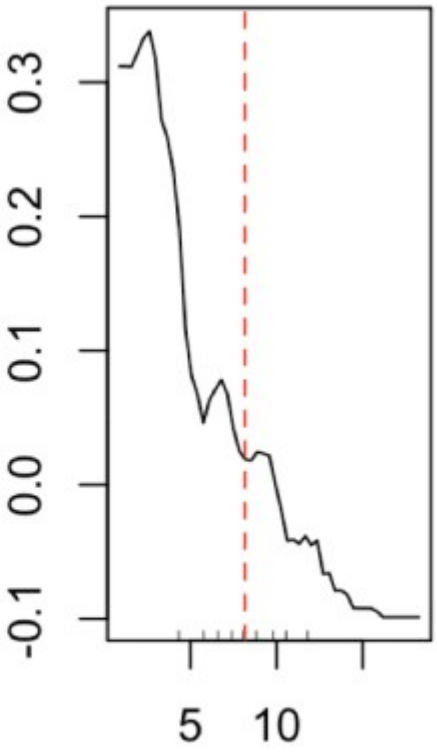
A



B



C



F

Predictions for the “average” point of the dataset is here

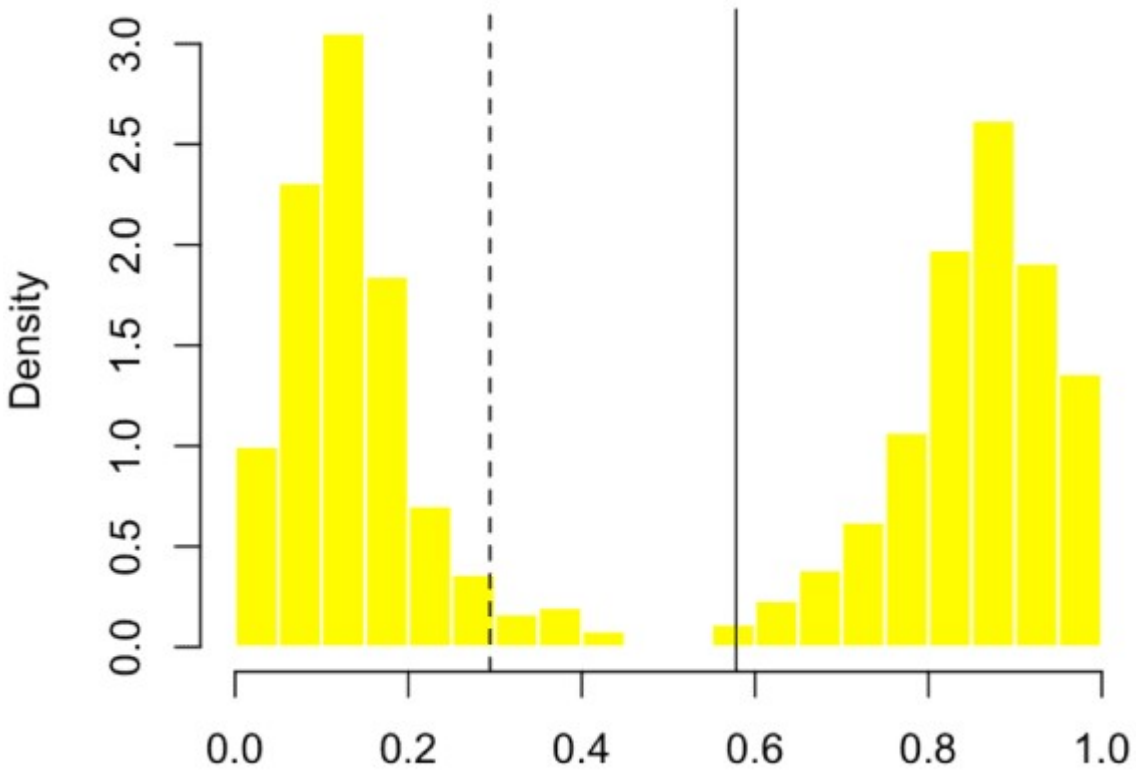
```
(parbre = predict(arbre,newdata=data.frame(t(apply(df[,-1],2,mean))
),type = "prob"))
```

```
       blue       red
1 0.8064516 0.1935484
(parbres = predict(arbres,newdata=data.frame(t(apply(df[,-1],2,mean))
),type = "prob"))
   blue   red
1 0.422 0.578
attr(,"class")
[1] "matrix" "votes"
```

and there is a substantial difference, with a probability of 19% with a single tree, 58% with 500 trees (the default value of the function).

To understand why we can have such a difference, we should not only focus on the bagging stratgy, but look at the variability of the predictions, obtained with trees,

```
B=1e4
parbres = rep(NA,B)
m=data.frame(t(apply(df[,-1],2,mean)))
for(b in 1:B){
  idx = sample(1:nrow(df),size=nrow(df),replace=TRUE)
  arbre = rpart(Y~., data=df[idx,])
  parbres[b] = predict(arbre,newdata=m,type = "prob")[2]
}
hist(parbres)
```



Surprisingly, we have here a bimodal function for $\hat{y}$ which is either very small for some trees, of very large for others. On average, we have a value close to 55%……