# Question 1:

Divide the saana.csv data randomly into two different datasets: ??? model calibration data (70%) ??? model evaluation data (30%) Build the models based on the calibration data and test the predictive performance of the models using the evaluation data. What is the predictive performance of the GLM, GAM and GBM models for Betnan, Cashyp, Empher and Salret based on AUC-values of the model evaluation data? Use mesotopo, soil\_moist, soil\_temp and soil\_ph as predictors. Report the results in one short paragraph (max 5 sentences). Note: you can use sample-function to divide the data, e.g.

#load data  
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING\_PHYSICAL\_GEOGRAPHY/assignment3/Data-20171114 (1)/saana.csv",sep=";")  
  
# Use the caTools package to extract the AUC values and compare them  
library(caTools)  
library(mgcv)  
library(gbm)

## Betnan

#number of times to repeat the models  
{rep<-10  
bet\_auc\_glm<-bet\_auc\_gam<-bet\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 #sample all the rows, and keep 70%(0.7)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,] #get the 70% rows for calibration  
 eva<- data[-rand\_sam,] #get the remaining 30% for evaluation  
   
 #create the glm for Betnan occurences  
 bet\_glm<-glm(Betnan~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="binomial")  
 #these could be used to select the Betnan but not necessary anymore. I used eva$Betnan instead  
 #eva\_bet<- eva[,grep("Betnan", colnames(data))]  
 #which(colnames(data)=="Betnan") or grep("Betnan", colnames(data))  
 pred\_bet\_glm<-predict.glm(bet\_glm, newdata = eva, type = "response")  
 #check the AUC of the compared prediction and evaluation  
 bet\_auc\_glm\_p<-colAUC(pred\_bet\_glm, eva$Betnan, plotROC=F)  
 bet\_auc\_glm <- c(bet\_auc\_glm, bet\_auc\_glm\_p[[1]])  
   
 #GAM  
 bet\_gam<-gam(Betnan~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="binomial")  
 pred\_bet\_gam<-predict.gam(bet\_gam, newdata = eva, type = "response")  
 bet\_auc\_gam\_p<-colAUC(pred\_bet\_gam, eva$Betnan, plotROC=F)  
 bet\_auc\_gam <- c(bet\_auc\_gam, bet\_auc\_gam\_p[[1]])  
   
 #GBM  
 bet\_gbm<-gbm(formula = Betnan~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(bet\_gbm, plot.it = F, method = "OOB")  
 pred\_bet\_gbm<-predict.gbm(bet\_gbm,newdata = eva, best.iter, type = "response")  
 bet\_auc\_gbm\_p<-colAUC(pred\_bet\_gbm, eva$Betnan, plotROC = F)  
 bet\_auc\_gbm<- c(bet\_auc\_gbm, bet\_auc\_gbm\_p[[1]])  
   
}   
compared\_model\_bat=cbind.data.frame(bet\_auc\_glm, bet\_auc\_gam, bet\_auc\_gbm)  
}  
#print the AUC values of the various models at different replications.  
compared\_model\_bat

## bet\_auc\_glm bet\_auc\_gam bet\_auc\_gbm  
## 1 0.5344828 0.5737548 0.6264368  
## 2 0.6180124 0.6925466 0.7142857  
## 3 0.5584291 0.6503831 0.6436782  
## 4 0.5710000 0.6990000 0.6990000  
## 5 0.5200846 0.6670190 0.6701903  
## 6 0.5009653 0.7084942 0.6515444  
## 7 0.6322222 0.7533333 0.7833333  
## 8 0.5138067 0.6084813 0.5113412  
## 9 0.5686499 0.7070938 0.6395881  
## 10 0.5771670 0.6310782 0.6522199

#mean of the the auc values for the various models  
colMeans(compared\_model\_bat)

## bet\_auc\_glm bet\_auc\_gam bet\_auc\_gbm   
## 0.5594820 0.6691184 0.6591618

#perform wilcoxon test to see if there is a significant improvement between the models  
wilcox.test(bet\_auc\_glm, bet\_auc\_gam, paired = T)

##   
## Wilcoxon signed rank test  
##   
## data: bet\_auc\_glm and bet\_auc\_gam  
## V = 0, p-value = 0.001953  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(bet\_auc\_gam, bet\_auc\_gbm, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: bet\_auc\_gam and bet\_auc\_gbm  
## V = 26, p-value = 0.7223  
## alternative hypothesis: true location shift is not equal to 0

\*\*\_ as it can be seen, there is a significant improvement from the glm to the gam. but insignificant from gam to gbm. However, with AUC below 0.8, the models are not strong enough.\_\*\*

## Cashyp

{rep<-7   
#create empty lists for the three models about to be built.  
 cas\_auc\_glm<-cas\_auc\_gam<-cas\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 #divide the sample into 70:30 training and testing respectively  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,] #calibration  
 eva<- data[-rand\_sam,] #evaluation  
   
 #create the prediction using glm  
 cas\_glm<-glm(Cashyp~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="binomial")  
 pred\_cas\_glm<-predict.glm(cas\_glm, newdata = eva, type = "response")  
 cas\_auc\_glm\_p<-colAUC(pred\_cas\_glm, eva$Cashyp, plotROC=F)  
 cas\_auc\_glm <- c(cas\_auc\_glm, cas\_auc\_glm\_p[[1]])  
   
 #GAM  
 cas\_gam<-gam(Cashyp~s(mesotopo, k=4) + s(soil\_moist, k=4) + s(soil\_temp, k=4) +   
 s(soil\_ph, k=4), data=cal,family ="binomial")  
 pred\_cas\_gam<-predict.gam(cas\_gam, newdata = eva, type = "response")  
 cas\_auc\_gam\_p<-colAUC(pred\_cas\_gam, eva$Cashyp, plotROC=F)  
 cas\_auc\_gam <- c(cas\_auc\_gam, cas\_auc\_gam\_p[[1]])  
   
 #GBM  
 cas\_gbm<-gbm(formula = Cashyp~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(cas\_gbm, plot.it = F, method = "OOB")  
 pred\_cas\_gbm<-predict.gbm(cas\_gbm,newdata = eva, best.iter, type = "response")  
 cas\_auc\_gbm\_p<-colAUC(pred\_cas\_gbm, eva$Cashyp, plotROC = F)  
 cas\_auc\_gbm<- c(cas\_auc\_gbm, cas\_auc\_gbm\_p[[1]])  
   
   
 }   
 compared\_model\_cas=cbind.data.frame(cas\_auc\_glm, cas\_auc\_gam, cas\_auc\_gbm)  
}  
#print the AUC values of the various models at different replications.  
compared\_model\_cas

## cas\_auc\_glm cas\_auc\_gam cas\_auc\_gbm  
## 1 0.8933333 0.8933333 0.8933333  
## 2 0.9378531 0.9378531 0.8389831  
## 3 0.9033333 0.7966667 0.8300000  
## 4 0.9566667 0.9533333 0.9466667  
## 5 0.9533333 0.9533333 0.9166667  
## 6 0.8446328 0.8446328 0.7994350  
## 7 0.9549180 0.9549180 0.8975410

#mean of the the auc values for the various models  
colMeans(compared\_model\_cas)

## cas\_auc\_glm cas\_auc\_gam cas\_auc\_gbm   
## 0.9205815 0.9048672 0.8746608

#perform wilcoxon test to see if there is a significant improvement between the models  
wilcox.test(cas\_auc\_glm, cas\_auc\_gam, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: cas\_auc\_glm and cas\_auc\_gam  
## V = 3, p-value = 0.3711  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(cas\_auc\_gam, cas\_auc\_gbm, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: cas\_auc\_gam and cas\_auc\_gbm  
## V = 19, p-value = 0.09349  
## alternative hypothesis: true location shift is not equal to 0

***Overall, the three models seem to be quite good without much difference, based on the parameters used.***

## Empher

{rep<-10  
#create empty lists to imput the models auc values later  
emp\_auc\_glm<-emp\_auc\_gam<-emp\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,]  
 eva<- data[-rand\_sam,]  
 emp\_glm<-glm(Empher~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="binomial")  
 pred\_emp\_glm<-predict.glm(emp\_glm, newdata = eva, type = "response")  
 emp\_auc\_glm\_p<-colAUC(pred\_emp\_glm, eva$Empher, plotROC=F)  
 emp\_auc\_glm <- c(emp\_auc\_glm, emp\_auc\_glm\_p[[1]])  
   
 #GAM  
 emp\_gam<-gam(Empher~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="binomial")  
 pred\_emp\_gam<-predict.gam(emp\_gam, newdata = eva, type = "response")  
 emp\_auc\_gam\_p<-colAUC(pred\_emp\_gam, eva$Empher, plotROC=F)  
 emp\_auc\_gam <- c(emp\_auc\_gam, emp\_auc\_gam\_p[[1]])  
   
 #GBM  
 emp\_gbm<-gbm(formula = Empher~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(emp\_gbm, plot.it = F, method = "OOB")  
 pred\_emp\_gbm<-predict.gbm(emp\_gbm,newdata = eva, best.iter, type = "response")  
 emp\_auc\_gbm\_p<-colAUC(pred\_emp\_gbm, eva$Empher, plotROC = F)  
 emp\_auc\_gbm<- c(emp\_auc\_gbm, emp\_auc\_gbm\_p[[1]])  
}   
#put all the results into a dataframe  
compared\_model\_emp=cbind.data.frame(emp\_auc\_glm, emp\_auc\_gam, emp\_auc\_gbm)  
}  
#show the results  
compared\_model\_emp

## emp\_auc\_glm emp\_auc\_gam emp\_auc\_gbm  
## 1 0.7966102 0.8079096 0.7655367  
## 2 0.5158730 0.5793651 0.6230159  
## 3 0.6106443 0.6344538 0.6960784  
## 4 0.7895623 0.6952862 0.7053872  
## 5 0.7828283 0.7508418 0.7474747  
## 6 0.6626667 0.6973333 0.6466667  
## 7 0.5476190 0.5826331 0.6778711  
## 8 0.6290909 0.5818182 0.6509091  
## 9 0.6430976 0.6767677 0.6801347  
## 10 0.7563636 0.6690909 0.7454545

#the mean of the aucs values of the models  
colMeans(compared\_model\_emp)

## emp\_auc\_glm emp\_auc\_gam emp\_auc\_gbm   
## 0.6734356 0.6675500 0.6938529

#compare the models  
wilcox.test(emp\_auc\_glm, emp\_auc\_gam, paired = T)

##   
## Wilcoxon signed rank test  
##   
## data: emp\_auc\_glm and emp\_auc\_gam  
## V = 29, p-value = 0.9219  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(emp\_auc\_gam, emp\_auc\_gbm, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: emp\_auc\_gam and emp\_auc\_gbm  
## V = 11.5, p-value = 0.1139  
## alternative hypothesis: true location shift is not equal to 0

**The gbm seems to be the best for Empher. However, all the models are not strong enough**

## Salret

{rep<-7 #number of replicatons of the sampling for the modelling  
#create empoty lists´to imput the auc values later  
sal\_auc\_glm<-sal\_auc\_gam<-sal\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,]  
 eva<- data[-rand\_sam,]  
 sal\_glm<-glm(Salret~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="binomial")  
 pred\_sal\_glm<-predict.glm(sal\_glm, newdata = eva, type = "response")  
 sal\_auc\_glm\_p<-colAUC(pred\_sal\_glm, eva$Salret, plotROC=F)  
 sal\_auc\_glm <- c(sal\_auc\_glm, sal\_auc\_glm\_p[[1]])  
   
 #GAM  
 sal\_gam<-gam(Salret~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="binomial")  
 pred\_sal\_gam<-predict.gam(sal\_gam, newdata = eva, type = "response")  
 sal\_auc\_gam\_p<-colAUC(pred\_sal\_gam, eva$Salret, plotROC=F)  
 sal\_auc\_gam <- c(sal\_auc\_gam, sal\_auc\_gam\_p[[1]])  
   
 #GBM  
 sal\_gbm<-gbm(formula = Salret~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(sal\_gbm, plot.it = F, method = "OOB")  
 pred\_sal\_gbm<-predict.gbm(sal\_gbm,newdata = eva, best.iter, type = "response")  
 sal\_auc\_gbm\_p<-colAUC(pred\_sal\_gbm, eva$Salret, plotROC = F)  
 sal\_auc\_gbm<- c(sal\_auc\_gbm, sal\_auc\_gbm\_p[[1]])  
}   
compared\_model\_sal=cbind.data.frame(sal\_auc\_glm, sal\_auc\_gam, sal\_auc\_gbm)  
}  
#show the results  
compared\_model\_sal

## sal\_auc\_glm sal\_auc\_gam sal\_auc\_gbm  
## 1 1.0000000 0.9892473 0.9784946  
## 2 0.9523810 0.9523810 0.9523810  
## 3 0.9126984 0.9285714 0.9761905  
## 4 0.8968254 0.8968254 0.8650794  
## 5 0.9761905 0.9761905 0.9206349  
## 6 0.9682540 0.9682540 0.9404762  
## 7 0.8933333 0.9066667 0.8466667

#the mean of the aucs values of the models  
colMeans(compared\_model\_sal)

## sal\_auc\_glm sal\_auc\_gam sal\_auc\_gbm   
## 0.9428118 0.9454480 0.9257033

#compare the models  
wilcox.test(sal\_auc\_glm, sal\_auc\_gam, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: sal\_auc\_glm and sal\_auc\_gam  
## V = 1, p-value = 0.4227  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(sal\_auc\_gam, sal\_auc\_gbm, paired = T)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: sal\_auc\_gam and sal\_auc\_gbm  
## V = 17, p-value = 0.2084  
## alternative hypothesis: true location shift is not equal to 0

**all the models for Salret seem strong, with glm performing best**

# Question 2.

What is the predictive performance of the GLM, GAM and GBM models for veg\_height and vasc\_spr? Again, build the models using calibration data and test the models using evaluation data. Use Spearman correlation -values as the evaluation metrics. Use the same set of predictors that you used in question 1). Report the results in one short paragraph (max 5 sentences).

data<- read.csv("C:/Users/oyeda/Desktop/MODELLING\_PHYSICAL\_GEOGRAPHY/assignment3/Data-20171114 (1)/saana.csv" ,sep=";")  
# Use the caTools package to extract the AUC values and compare them  
#library(caTools)  
#library(mgcv)  
#library(gbm)  
  
#number of times to repeat the models  
{rep<-10  
 h\_auc\_glm<-auc\_gam<-h\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 #sample all the rows, and keep 70%(0.7)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,] #get the 70% rows for calibration  
 eva<- data[-rand\_sam,] #get the remaining 30% for evaluation  
   
 #create the glm for veg\_height occurences  
 h\_glm<-glm(veg\_height~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="gaussian")  
 pred\_h\_glm<-predict.glm(h\_glm, newdata = eva, type = "response")  
 h\_cor\_glm<-cor(pred\_h\_glm, eva$veg\_height, method = "spearman")  
   
   
 #GAM  
 h\_gam<-gam(veg\_height~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="gaussian")  
 pred\_h\_gam<-predict.gam(h\_gam, newdata = eva, type = "response")  
 h\_cor\_gam<-cor(pred\_h\_gam, eva$veg\_height, method = "spearman")  
   
 #GBM  
 h\_gbm<-gbm(formula = veg\_height~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=data,  
 distribution = "gaussian",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(h\_gbm, plot.it = F, method = "OOB")  
 pred\_h\_gbm<-predict.gbm(h\_gbm,newdata = eva, best.iter, type = "response")  
 h\_cor\_gbm<-cor(pred\_h\_gbm, eva$veg\_height, method = "spearman")  
 }   
 compared\_model\_h=cbind.data.frame(h\_cor\_glm, h\_cor\_gam, h\_cor\_gbm)  
}  
#comparison between the corelation between predicted and observed vegetation height of the models  
compared\_model\_h

## h\_cor\_glm h\_cor\_gam h\_cor\_gbm  
## 1 0.397975 0.4106639 0.4999944

***Using the three models, they show low correlation with the observed vegetation height from the evaluation/testing data***

## vasc\_spr

{rep<-7  
vspr\_auc\_glm<-vspr\_auc\_gam<-vspr\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 #sample all the rows, and keep 70%(0.7)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,] #get the 70% rows for calibration  
 eva<- data[-rand\_sam,] #get the remaining 30% for evaluation  
   
 #create the glm for veg\_height occurences  
 vspr\_glm<-glm(vasc\_spr~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="poisson")  
 pred\_vspr\_glm<-predict.glm(vspr\_glm, newdata = eva, type = "response")  
 vspr\_cor\_glm<-cor(pred\_vspr\_glm, eva$vasc\_spr, method = "spearman")  
   
 #GAM  
 vspr\_gam<-gam(vasc\_spr~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="poisson")  
 pred\_vspr\_gam<-predict.gam(vspr\_gam, newdata = eva, type = "response")  
 vspr\_cor\_gam<-cor(pred\_vspr\_gam, eva$vasc\_spr, method = "spearman")  
   
 #GBM  
 vspr\_gbm<-gbm(formula = vasc\_spr~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=data,  
 distribution = "poisson",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(vspr\_gbm, plot.it = F, method = "OOB")  
 pred\_vspr\_gbm<-predict.gbm(vspr\_gbm,newdata = eva, best.iter, type = "response")  
 vspr\_cor\_gbm<-cor(pred\_vspr\_gbm, eva$vasc\_spr, method = "spearman")  
}   
compared\_model\_vspr=cbind.data.frame(vspr\_cor\_glm, vspr\_cor\_gam, vspr\_cor\_gbm)  
}  
#compare the models  
compared\_model\_vspr

## vspr\_cor\_glm vspr\_cor\_gam vspr\_cor\_gbm  
## 1 0.7181524 0.7355887 0.8446751

***Here, the GBM seems to be the best and others are fairly good too but less reliable***

# Question 3.

Characterize soil\_moist, soil\_temp, soil\_ph, veg\_height and vasc\_spr conditions along the mesotopographic gradient using GAM. Model the values of these five responses at the valley bottom (mesotopo 1), mid-slope (mesotopo 5) and ridge-top (mesotopo 10). Present the results as an informative figure. Report the results in one short paragraph (max 5 sentences).

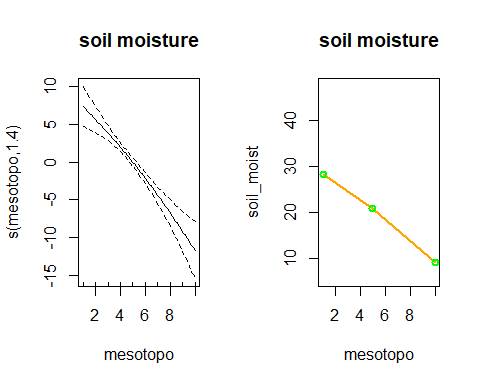
#library(mgcv)  
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING\_PHYSICAL\_GEOGRAPHY/assignment3/Data-20171114 (1)/saana.csv"  
 ,sep=";")  
attach(data)  
#This is the first method makes the prediction without classifying the mesotopography  
#GAM

### Soil moisture

gam\_moist <- gam(soil\_moist~s(mesotopo, k=3), data = data, family = "gaussian")  
summary(gam\_moist)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## soil\_moist ~ s(mesotopo, k = 3)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.7514 0.4879 42.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(mesotopo) 1.396 1.635 36.71 3.9e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.228 Deviance explained = 23.3%  
## GCV = 51.999 Scale est. = 51.423 n = 216

par(mfrow=c(1,2))  
plot(gam\_moist, main = "soil moisture")  
  
#the values at the valley bottom, mid-slope and ridge-top  
mesotopo2 <- c(1,5,10)  
newdata <- data.frame(mesotopo=mesotopo2)  
pred.gam\_moist <- predict.gam(gam\_moist, newdata, type="response")  
plot(mesotopo, soil\_moist, pch=19, cex=0.2, col="grey",type="n", main="soil moisture")   
lines(mesotopo2, pred.gam\_moist, lty=1,lwd= 2,col="orange")   
points(mesotopo2, pred.gam\_moist, lty=1,lwd= 2,col="green")

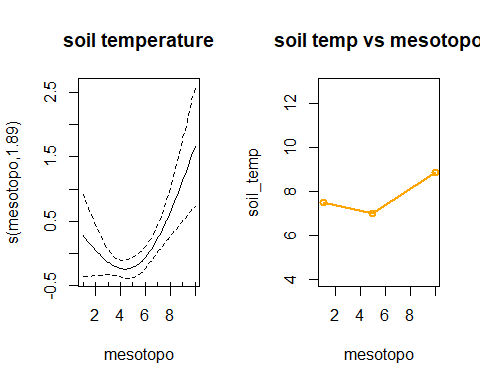
 Overall, soil moisture seems to be reducing with increase in mesotopographical gradient. The response curve shows high confidence at the mid-slope which shos that there are more samples from tha area.

### soil temperature

gam\_temp<- gam(soil\_temp~s(mesotopo, k=3), data = data, family = "gaussian")  
summary(gam\_temp) #summary soil temperature

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## soil\_temp ~ s(mesotopo, k = 3)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.21135 0.09504 75.88 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(mesotopo) 1.889 1.988 5.965 0.00216 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.0526 Deviance explained = 6.09%  
## GCV = 1.9773 Scale est. = 1.9509 n = 216

par(mfrow=c(1,2))  
plot(gam\_temp, main="soil temperature") #response curve  
pred.gam\_temp <- predict.gam(gam\_temp, newdata, type="response")  
plot(mesotopo, soil\_temp, pch=19, cex=0.2, col="grey",type="n", main = "soil temp vs mesotopo")   
lines(mesotopo2, pred.gam\_temp, lty=1,lwd= 2,col="orange")  
points(mesotopo2, pred.gam\_temp, lty=1,lwd= 2,col="orange")

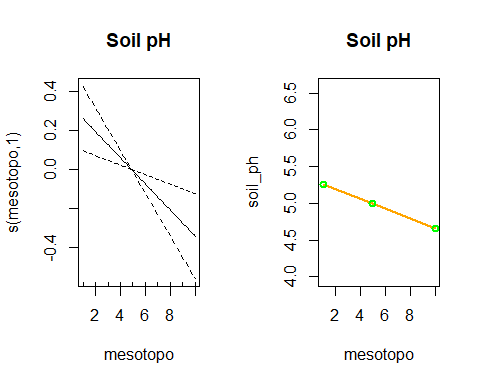
 ***soil temperature seems to be slightly reducing as one approaches mid-slope, then increases towards the rigde-top. This could perhaps,be because there are lesser/sparse vegeation at the ridge-top, thereby exposing the soil to solar radiation. However, the confidence are slower at the valley-bottom and ridge-top, ostensibly because, there are less data from those areas.***

### soil pH

gam\_ph<- gam(soil\_ph~s(mesotopo, k=3), data = data, family = "gaussian")  
summary(gam\_ph) #summary soil\_pH

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## soil\_ph ~ s(mesotopo, k = 3)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.99978 0.04018 124.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(mesotopo) 1 1 9.893 0.00189 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.0397 Deviance explained = 4.42%  
## GCV = 0.35205 Scale est. = 0.34879 n = 216

par(mfrow=c(1,2))  
plot(gam\_ph, main = "Soil pH") #response curve  
  
pred.gam\_ph <- predict.gam(gam\_ph, newdata, type="response")  
plot(mesotopo, soil\_ph, pch=19, cex=0.2, col="grey",type="n", main = "Soil pH")   
lines(mesotopo2, pred.gam\_ph, lty=1,lwd= 2,col="orange")  
points(mesotopo2, pred.gam\_ph, lty=1,lwd= 2,col="green")

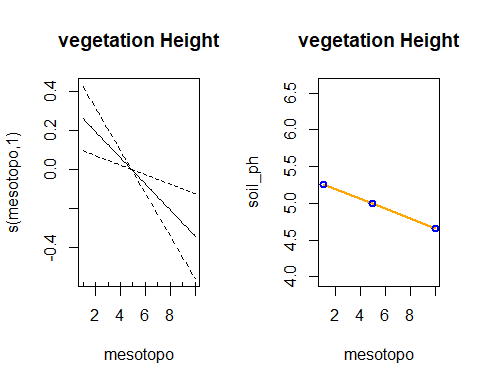
 The soil ph also appears to be reducing upslope. This means that the soil upslope are more acidic. However, the confidence levels are low at the valley bottom and ridge-top

### Vegetation Height

par(mfrow=c(1,2))  
gam\_vh<- gam(veg\_height~s(mesotopo, k=3), data = data, family = "poisson")  
summary(gam\_vh)

##   
## Family: poisson   
## Link function: log   
##   
## Formula:  
## veg\_height ~ s(mesotopo, k = 3)  
##   
## Parametric coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.70664 0.02929 58.27 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df Chi.sq p-value   
## s(mesotopo) 1.87 1.983 34.16 2.29e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.0724 Deviance explained = 10.4%  
## UBRE = 0.6694 Scale est. = 1 n = 216

plot(gam\_ph, main = "vegetation Height")  
pred.gam\_ph <- predict.gam(gam\_ph, newdata, type="response")  
plot(mesotopo, soil\_ph, pch=19, cex=0.2, col="grey",type="n", main = "vegetation Height")   
lines(mesotopo2, pred.gam\_ph, lty=1,lwd= 2,col="orange")  
points(mesotopo2, pred.gam\_ph, lty=1,lwd= 2,col="blue")

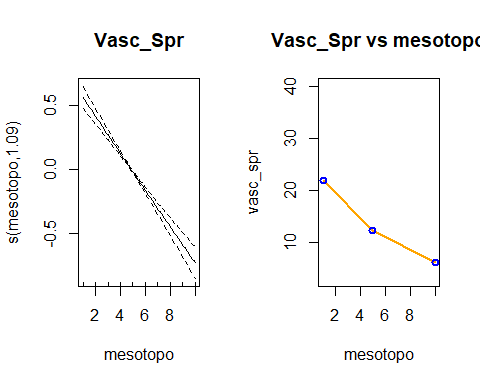
 ***This is also similar to the soil pH. This is expected, as height would be affected by the acidity and soil moisture which is lower upslope. The vegetation height seems to be reducing, as one approaches upslope***

### Vascular species richness

gam\_vaspr<- gam(vasc\_spr~s( mesotopo, k=3), data = data, family = "poisson")  
summary(gam\_vaspr)

##   
## Family: poisson   
## Link function: log   
##   
## Formula:  
## vasc\_spr ~ s(mesotopo, k = 3)  
##   
## Parametric coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.52237 0.01962 128.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df Chi.sq p-value   
## s(mesotopo) 1.085 1.163 189 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.248 Deviance explained = 27%  
## UBRE = 1.4259 Scale est. = 1 n = 216

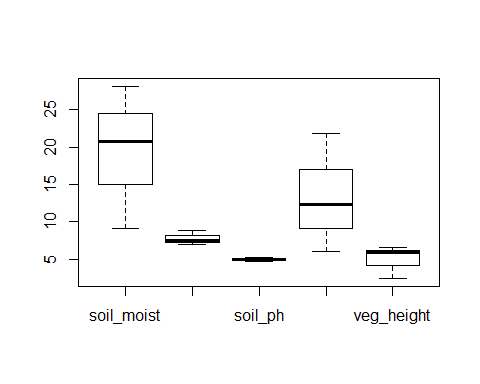
par(mfrow=c(1,2))  
plot(gam\_vaspr, main = "Vasc\_Spr")  
pred.gam\_vaspr <- predict.gam(gam\_vaspr, newdata, type="response")  
plot(mesotopo, vasc\_spr, pch=19, cex=0.2, col="grey",type="n", main = "Vasc\_Spr vs mesotopo")   
lines(mesotopo2, pred.gam\_vaspr, lty=1,lwd= 2,col="orange")  
points(mesotopo2, pred.gam\_vaspr, lty=1,lwd= 2,col="blue")

 ***vascular species richness also seems tobe reducing uplspe. Same conditions that affect vegeatin heights are expected to affect the vascular species richness.***

###############################  
#data$topo\_level<-cut(mesotopo, breaks = c(0,4,7,10))  
#levels(data$topo\_level)<-c("valley-bottom","mid-slope","ridge-top")  
  
#Subsetting the mesotop into the various parts  
val\_bot<-data[mesotopo==1,] #valley bottom  
mid\_sl<-data[mesotopo==5,] #mid-slope  
r\_top<-data[mesotopo==10,] #ridge-top  
  
  
#create data frame to impute the modlled values at variouis topo gradients.  
topo<-matrix(ncol = 5, nrow = 3)  
topo<- data.frame(topo)  
row.names(topo)<- c("valley\_bottom", "mid-slope", "ridge-top")  
colnames(topo)<-c("soil\_moist", "soil\_temp", "soil\_ph", "vasc\_spr", "veg\_height")  
  
  
#predicting the values at the valley bottom for the responses  
vb1<- topo[1,1]<- mean(predict.gam(gam\_moist, val\_bot, type="response"))  
vb2<- topo[1,2]<- mean(predict.gam(gam\_temp, val\_bot, type="response"))  
vb3<- topo[1,3]<- mean(predict.gam(gam\_ph, val\_bot, type="response"))  
vb4<- topo[1,4]<- mean(predict.gam(gam\_vaspr, val\_bot, type="response"))  
vb5<- topo[1,5]<- mean(predict.gam(gam\_vh, val\_bot, type="response"))  
  
#predicting the values at the mid-slope for the responses  
ms1<- topo[2,1]<- mean(predict.gam(gam\_moist, mid\_sl, type="response"))  
ms2<- topo[2,2]<- mean(predict.gam(gam\_temp, mid\_sl, type="response"))  
ms3<- topo[2,3]<- mean(predict.gam(gam\_ph, mid\_sl, type="response"))  
ms4<- topo[2,4]<- mean(predict.gam(gam\_vaspr, mid\_sl, type="response"))  
ms5<- topo[2,5]<- mean(predict.gam(gam\_vh, mid\_sl, type="response"))  
  
#predicting the values at the ridge-top for the responses  
rt1<- topo[3,1]<- mean(predict.gam(gam\_moist, r\_top, type="response"))  
rt2<- topo[3,2]<- mean(predict.gam(gam\_temp, r\_top, type="response"))  
rt3<- topo[3,3]<- mean(predict.gam(gam\_ph, r\_top, type="response"))  
rt4<- topo[3,4]<- mean(predict.gam(gam\_vaspr, r\_top, type="response"))  
rt5<- topo[3,5]<- mean(predict.gam(gam\_vh, r\_top, type="response"))  
  
#view the dataframe  
topo

## soil\_moist soil\_temp soil\_ph vasc\_spr veg\_height  
## valley\_bottom 28.124142 7.487852 5.261009 21.851141 6.556635  
## mid-slope 20.752299 6.996628 4.992635 12.250299 5.830233  
## ridge-top 9.100299 8.875371 4.657167 6.047608 2.411296

#see the boxplot  
boxplot(topo)

 There seems to be trends in variables changes across slope

# Question 4.

Does the cover of Empetrum hermaphroditum (Empher\_cover) have an effect on the vasc\_spr when all other predictors are controlled for? Use the same set of predictors as used in question 1). Use all three modelling frameworks to test the hypothesis. Report the results in one short paragraph (max 5 sentences).The main idea behind this question: as a dominant species Empher\_cover might have a strong influence on the vegetation properties - can we see the effect? Please, test it!

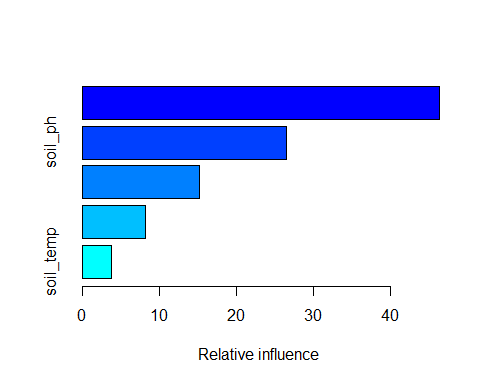
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING\_PHYSICAL\_GEOGRAPHY/assignment3/Data-20171114 (1)/saana.csv",sep=";")  
  
# Use the caTools package to extract the AUC values and compare them  
#library(caTools)  
#library(mgcv)  
#library(gbm)  
  
attach(data)  
  
#create the glm for vasc\_spr occurences'  
#GLM  
vaspr\_glm<-glm(vasc\_spr~Empher\_cover+mesotopo+soil\_moist+soil\_temp+soil\_ph, data=data,family ="poisson")  
summary(vaspr\_glm)

##   
## Call:  
## glm(formula = vasc\_spr ~ Empher\_cover + mesotopo + soil\_moist +   
## soil\_temp + soil\_ph, family = "poisson", data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9517 -0.8621 -0.1215 0.7040 2.8930   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.638905 0.267599 2.388 0.017 \*   
## Empher\_cover -0.005616 0.001012 -5.552 2.82e-08 \*\*\*  
## mesotopo -0.053376 0.012330 -4.329 1.50e-05 \*\*\*  
## soil\_moist 0.025377 0.002470 10.275 < 2e-16 \*\*\*  
## soil\_temp 0.016219 0.015725 1.031 0.302   
## soil\_ph 0.310197 0.034137 9.087 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 712.23 on 215 degrees of freedom  
## Residual deviance: 241.45 on 210 degrees of freedom  
## AIC: 1178.1  
##   
## Number of Fisher Scoring iterations: 4

#GAM  
vaspr\_gam<-gam(vasc\_spr~s(Empher\_cover)+s(mesotopo)+s(soil\_moist)+  
 s(soil\_temp)+s(soil\_ph), data=data,family ="poisson")  
summary(vaspr\_gam)

##   
## Family: poisson   
## Link function: log   
##   
## Formula:  
## vasc\_spr ~ s(Empher\_cover) + s(mesotopo) + s(soil\_moist) + s(soil\_temp) +   
## s(soil\_ph)  
##   
## Parametric coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.47080 0.02048 120.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df Chi.sq p-value   
## s(Empher\_cover) 5.270 6.387 55.081 1.11e-09 \*\*\*  
## s(mesotopo) 2.669 3.387 10.699 0.0172 \*   
## s(soil\_moist) 1.744 2.185 91.887 < 2e-16 \*\*\*  
## s(soil\_temp) 1.045 1.089 1.415 0.2707   
## s(soil\_ph) 2.687 3.363 97.433 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.735 Deviance explained = 73.3%  
## UBRE = 0.013044 Scale est. = 1 n = 216

#GBM  
vaspr\_gbm<-gbm(formula = vasc\_spr~Empher\_cover+mesotopo+soil\_moist+soil\_temp+soil\_ph, data=data,  
 distribution = "poisson",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
summary(vaspr\_gbm)



## var rel.inf  
## soil\_moist soil\_moist 46.312510  
## soil\_ph soil\_ph 26.451865  
## Empher\_cover Empher\_cover 15.225292  
## mesotopo mesotopo 8.239242  
## soil\_temp soil\_temp 3.771091

***From the above, Empher\_cover appears to be a significant predictor and has the third relative importance as shown in GBM. Therefore, we can say Empher\_Cover has a strong influence on vegetation properties***

### testing the model Without Empher\_Cover

{rep<-10  
 vaspr\_auc\_glm<-vaspr\_auc\_gam<-vaspr\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,]  
 eva<- data[-rand\_sam,]  
 vaspr\_glm<-glm(vasc\_spr~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="poisson")  
 pred\_vaspr\_glm<-predict.glm(vaspr\_glm, newdata = eva, type = "response")  
 vaspr\_auc\_glm\_p<-colAUC(pred\_vaspr\_glm, eva$vasc\_spr, plotROC=F)  
 vaspr\_auc\_glm <- c(vaspr\_auc\_glm, vaspr\_auc\_glm\_p[[1]])  
   
 #GAM  
 vaspr\_gam<-gam(vasc\_spr~s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="poisson")  
 pred\_vaspr\_gam<-predict.gam(vaspr\_gam, newdata = eva, type = "response")  
 vaspr\_auc\_gam\_p<-colAUC(pred\_vaspr\_gam, eva$vasc\_spr, plotROC=F)  
 vaspr\_auc\_gam <- c(vaspr\_auc\_gam, vaspr\_auc\_gam\_p[[1]])  
   
 #GBM  
 vaspr\_gbm<-gbm(formula = vasc\_spr~mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "poisson",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(vaspr\_gbm, plot.it = F, method = "OOB")  
 pred\_vaspr\_gbm<-predict.gbm(vaspr\_gbm,newdata = eva, best.iter, type = "response")  
 vaspr\_auc\_gbm\_p<-colAUC(pred\_vaspr\_gbm, eva$vasc\_spr, plotROC = F)  
 vaspr\_auc\_gbm<- c(vaspr\_auc\_gbm, vaspr\_auc\_gbm\_p[[1]])  
 }   
 compared\_model\_vaspr1=cbind.data.frame(vaspr\_auc\_glm, vaspr\_auc\_gam, vaspr\_auc\_gbm)  
}

### Tesing the model when Empher\_cover is included in the prediction

{rep<-10  
 vaspr\_auc\_glm<-vaspr\_auc\_gam<-vaspr\_auc\_gbm<-c()  
for (i in 1:rep){  
 #print(i)  
 rand\_sam<-sample(1:nrow(data), size = 0.7\*nrow(data) )  
 cal<- data[rand\_sam,]  
 eva<- data[-rand\_sam,]  
 vaspr\_glm<-glm(vasc\_spr~Empher\_cover+mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,family ="poisson")  
 pred\_vaspr\_glm<-predict.glm(vaspr\_glm, newdata = eva, type = "response")  
 vaspr\_auc\_glm\_p<-colAUC(pred\_vaspr\_glm, eva$vasc\_spr, plotROC=F)  
 vaspr\_auc\_glm <- c(vaspr\_auc\_glm, vaspr\_auc\_glm\_p[[1]])  
   
 #GAM  
 vaspr\_gam<-gam(vasc\_spr~s(Empher\_cover, k=3)+ s(mesotopo, k=3) + s(soil\_moist, k=3) + s(soil\_temp, k=3) +   
 s(soil\_ph, k=3), data=cal,family ="poisson")  
 pred\_vaspr\_gam<-predict.gam(vaspr\_gam, newdata = eva, type = "response")  
 vaspr\_auc\_gam\_p<-colAUC(pred\_vaspr\_gam, eva$vasc\_spr, plotROC=F)  
 vaspr\_auc\_gam <- c(vaspr\_auc\_gam, vaspr\_auc\_gam\_p[[1]])  
   
 #GBM  
 vaspr\_gbm<-gbm(formula = vasc\_spr~Empher\_cover+mesotopo+soil\_moist+soil\_temp+soil\_ph, data=cal,  
 distribution = "poisson",n.trees = 3000, shrinkage = 0.001, interaction.depth = 4)  
 best.iter<-gbm.perf(vaspr\_gbm, plot.it = F, method = "OOB")  
 pred\_vaspr\_gbm<-predict.gbm(vaspr\_gbm,newdata = eva, best.iter, type = "response")  
 vaspr\_auc\_gbm\_p<-colAUC(pred\_vaspr\_gbm, eva$vasc\_spr, plotROC = F)  
 vaspr\_auc\_gbm<- c(vaspr\_auc\_gbm, vaspr\_auc\_gbm\_p[[1]])  
 }   
 compared\_model\_vaspr2=cbind.data.frame(vaspr\_auc\_glm, vaspr\_auc\_gam, vaspr\_auc\_gbm)  
}  
  
compared\_model\_vaspr1

## vaspr\_auc\_glm vaspr\_auc\_gam vaspr\_auc\_gbm  
## 1 1.0000000 1.0000000 1.0000000  
## 2 0.6000000 0.7000000 0.6000000  
## 3 0.8333333 0.8333333 0.6666667  
## 4 0.8000000 0.8000000 1.0000000  
## 5 0.7500000 1.0000000 0.5000000  
## 6 1.0000000 1.0000000 0.5000000  
## 7 1.0000000 0.5000000 1.0000000  
## 8 0.5000000 1.0000000 1.0000000  
## 9 0.7500000 0.5833333 0.7500000  
## 10 0.5833333 0.6666667 0.6666667

compared\_model\_vaspr2

## vaspr\_auc\_glm vaspr\_auc\_gam vaspr\_auc\_gbm  
## 1 1.0000000 0.6666667 1.0000000  
## 2 0.7500000 0.7500000 0.7500000  
## 3 0.5000000 1.0000000 1.0000000  
## 4 0.7500000 0.6250000 0.8750000  
## 5 0.5000000 0.5000000 1.0000000  
## 6 0.5000000 0.7500000 1.0000000  
## 7 0.6666667 0.6666667 0.6666667  
## 8 1.0000000 0.5000000 1.0000000  
## 9 1.0000000 1.0000000 1.0000000  
## 10 0.5000000 0.5000000 0.6666667

wilcox.test(mean(compared\_model\_vaspr1[,1]), mean(compared\_model\_vaspr2[,1]))

##   
## Wilcoxon rank sum test  
##   
## data: mean(compared\_model\_vaspr1[, 1]) and mean(compared\_model\_vaspr2[, 1])  
## W = 1, p-value = 1  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(mean(compared\_model\_vaspr1[,2]), mean(compared\_model\_vaspr2[,2]))

##   
## Wilcoxon rank sum test  
##   
## data: mean(compared\_model\_vaspr1[, 2]) and mean(compared\_model\_vaspr2[, 2])  
## W = 1, p-value = 1  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(mean(compared\_model\_vaspr1[,3]), mean(compared\_model\_vaspr2[,3]))

##   
## Wilcoxon rank sum test  
##   
## data: mean(compared\_model\_vaspr1[, 3]) and mean(compared\_model\_vaspr2[, 3])  
## W = 0, p-value = 1  
## alternative hypothesis: true location shift is not equal to 0

***I decided to try my hands on building the model with and without Empher\_cover and comparing the auc values. There seems to be a slight improvement, as shown in the tables, However, the predictions did not improve significantly***