Model

The Dark Knight

10/5/2020

```r
# load data
df.train <- read.csv("/Users/hoichunlaw/Documents/w210/data/train_data_with_clusters_DBSCAN.csv")
df.test <- read.csv("/Users/hoichunlaw/Documents/w210/data/test_data_with_clusters_DBSCAN.csv")
#names(df)
```

```r
# data manipulation
df.train$cluster_location <- factor(df.train$cluster_location)
df.train$cluster_weather <- factor(df.train$cluster_weather)
df.train$cluster_weather_DBSCAN <- factor(df.train$cluster_weather_DBSCAN)
df.train$PhyloClust56 <- factor(df.train$PhyloClust56)
df.train$AET_divided_by_PET <- df.train$X30.1_AET_Mean_mm / df.train$X30.2_PET_Mean_mm
df.train$log_poultry <- log(df.train$poultry)
df.train$log_livestock_mam <- log(df.train$livestock_mam)

df.test$cluster_location <- factor(df.test$cluster_location)
df.test$cluster_weather <- factor(df.test$cluster_weather)
df.test$cluster_weather_DBSCAN <- factor(df.test$cluster_weather_DBSCAN)
df.test$PhyloClust56 <- factor(df.test$PhyloClust56)
df.test$AET_divided_by_PET <- df.test$X30.1_AET_Mean_mm / df.test$X30.2_PET_Mean_mm
df.test$log_poultry <- log(df.test$poultry)
df.test$log_livestock_mam <- log(df.test$livestock_mam)
```

Build Poisson Regression with stepwise forward method base on AIC

```
# select feature set
features = c("X27.4_HuPopDen_Change", "cluster_weather_DBSCAN", "cluster_location",
             "X30.1_AET_Mean_mm", "X30.2_PET_Mean_mm",
             "AET_divided_by_PET", "earth2_trees_everg", "crop_change",
             "mamdiv", "earth11_barren",
             "log_poultry", "log_livestock_mam", "earth7_veg_manag", "PhyloClust56")

# select data with sample size > 50
df.train <- df.train[df.train$Total > 50,]
df.train$count <- round(df.train$Positive / df.train$Total * 100)

empty.mod <- glm(count ~ 1, family=poisson(link=log), data=df.train)
full.mod <- glm(count ~ ., family=poisson(link=log), data=df.train[,c(features, "count")])
forw.sel <- step(object=empty.mod, scope = list(upper=full.mod), direction="forward", k=log(nrow(df.trai
```

```
## Start:  AIC=923.42
## count ~ 1
##
##                          Df Deviance    AIC
## + PhyloClust56            5   524.65 830.07
## + mamdiv                  1   577.65 865.08
## + earth11_barren          1   609.83 897.26
## + earth2_trees_everg      1   614.09 901.51
## + AET_divided_by_PET      1   614.27 901.70
## + log_livestock_mam       1   617.74 905.16
## + earth7_veg_manag        1   618.93 906.36
## + X30.1_AET_Mean_mm       1   624.81 912.24
## + log_poultry             1   628.10 915.52
## + cluster_location        4   619.06 919.99
## + cluster_weather_DBSCAN  3   624.01 920.44
## + crop_change             1   633.83 921.26
## + X27.4_HuPopDen_Change   1   634.50 921.93
## <none>                        640.49 923.42
## + X30.2_PET_Mean_mm       1   639.68 927.11
##
## Step:  AIC=830.07
## count ~ PhyloClust56
##
##                          Df Deviance    AIC
## + earth7_veg_manag        1   499.37 809.30
## + mamdiv                  1   499.44 809.36
## + log_livestock_mam       1   499.99 809.91
## + log_poultry             1   517.43 827.35
## + earth11_barren          1   518.42 828.35
## <none>                        524.65 830.07
## + X30.2_PET_Mean_mm       1   522.33 832.26
## + X27.4_HuPopDen_Change   1   522.91 832.84
## + crop_change             1   523.38 833.31
## + AET_divided_by_PET      1   523.82 833.75
## + earth2_trees_everg      1   524.06 833.98
## + X30.1_AET_Mean_mm       1   524.50 834.43
## + cluster_weather_DBSCAN  3   515.76 834.69
```

```
## + cluster_location        4   516.81 840.24
##
## Step:  AIC=809.3
## count ~ PhyloClust56 + earth7_veg_manag
##
##                          Df Deviance    AIC
## + crop_change             1   481.66 796.09
## + mamdiv                  1   484.08 798.51
## + cluster_weather_DBSCAN  3   478.54 801.97
## + X27.4_HuPopDen_Change   1   490.02 804.44
## <none>                        499.37 809.30
## + cluster_location        4   482.72 810.64
## + AET_divided_by_PET      1   496.66 811.09
## + log_livestock_mam       1   496.72 811.14
## + earth2_trees_everg      1   497.37 811.80
## + log_poultry             1   497.52 811.95
## + X30.1_AET_Mean_mm       1   498.80 813.23
## + X30.2_PET_Mean_mm       1   499.24 813.66
## + earth11_barren          1   499.34 813.76
##
## Step:  AIC=796.09
## count ~ PhyloClust56 + earth7_veg_manag + crop_change
##
##                          Df Deviance    AIC
## + X30.2_PET_Mean_mm       1   458.95 777.88
## + X30.1_AET_Mean_mm       1   467.05 785.97
## + cluster_weather_DBSCAN  3   462.95 790.87
## <none>                        481.66 796.09
## + AET_divided_by_PET      1   477.20 796.12
## + mamdiv                  1   478.83 797.75
## + earth2_trees_everg      1   479.58 798.51
## + log_livestock_mam       1   479.97 798.89
## + log_poultry             1   480.10 799.02
## + earth11_barren          1   480.44 799.37
## + cluster_location        4   467.50 799.93
## + X27.4_HuPopDen_Change   1   481.41 800.34
##
## Step:  AIC=777.88
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm
##
##                          Df Deviance    AIC
## + cluster_weather_DBSCAN  3   429.54 761.96
## + cluster_location        4   433.75 770.68
## + log_livestock_mam       1   449.64 773.06
## + mamdiv                  1   450.70 774.13
## + X27.4_HuPopDen_Change   1   452.68 776.11
## <none>                        458.95 777.88
## + AET_divided_by_PET      1   457.06 780.48
## + earth2_trees_everg      1   458.26 781.69
## + X30.1_AET_Mean_mm       1   458.79 782.22
## + earth11_barren          1   458.88 782.31
## + log_poultry             1   458.91 782.34
##
## Step:  AIC=761.96
```

```
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN
##
##                          Df Deviance    AIC
## + log_livestock_mam      1   414.10 751.02
## + X27.4_HuPopDen_Change  1   415.44 752.37
## + mamdiv                 1   421.13 758.05
## <none>                       429.54 761.96
## + AET_divided_by_PET     1   425.63 762.55
## + log_poultry            1   426.12 763.04
## + cluster_location       4   413.78 764.21
## + earth2_trees_everg     1   429.06 765.99
## + X30.1_AET_Mean_mm      1   429.49 766.42
## + earth11_barren         1   429.51 766.43
##
## Step:  AIC=751.02
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam
##
##                          Df Deviance    AIC
## + X27.4_HuPopDen_Change  1   399.68 741.10
## + mamdiv                 1   405.85 747.28
## <none>                       414.10 751.02
## + cluster_location       4   398.89 753.81
## + log_poultry            1   412.42 753.85
## + AET_divided_by_PET     1   412.87 754.29
## + X30.1_AET_Mean_mm      1   413.38 754.80
## + earth2_trees_everg     1   413.86 755.29
## + earth11_barren         1   414.03 755.46
##
## Step:  AIC=741.1
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change
##
##                       Df Deviance    AIC
## + cluster_location    4   376.76 736.18
## <none>                    399.68 741.10
## + mamdiv              1   396.95 742.87
## + X30.1_AET_Mean_mm   1   397.52 743.44
## + log_poultry         1   398.19 744.11
## + earth2_trees_everg  1   398.64 744.56
## + AET_divided_by_PET  1   399.65 745.58
## + earth11_barren      1   399.67 745.59
##
## Step:  AIC=736.18
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change +
##     cluster_location
##
##                       Df Deviance    AIC
## + mamdiv              1   360.67 724.60
## <none>                    376.76 736.18
## + log_poultry         1   372.72 736.64
## + X30.1_AET_Mean_mm   1   373.66 737.59
```

```
## + earth2_trees_everg  1   374.34 738.27
## + earth11_barren      1   375.59 739.52
## + AET_divided_by_PET  1   376.64 740.57
##
## Step:  AIC=724.6
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change +
##     cluster_location + mamdiv
##
##                      Df Deviance    AIC
## + earth11_barren      1   350.36 718.78
## + earth2_trees_everg  1   354.73 723.15
## <none>                    360.67 724.60
## + log_poultry         1   359.57 728.00
## + AET_divided_by_PET  1   359.76 728.18
## + X30.1_AET_Mean_mm   1   360.49 728.91
##
## Step:  AIC=718.78
## count ~ PhyloClust56 + earth7_veg_manag + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change +
##     cluster_location + mamdiv + earth11_barren
##
##                      Df Deviance    AIC
## <none>                    350.36 718.78
## + earth2_trees_everg  1   348.06 720.99
## + log_poultry         1   348.24 721.17
## + X30.1_AET_Mean_mm   1   348.46 721.39
## + AET_divided_by_PET  1   350.34 723.26
```

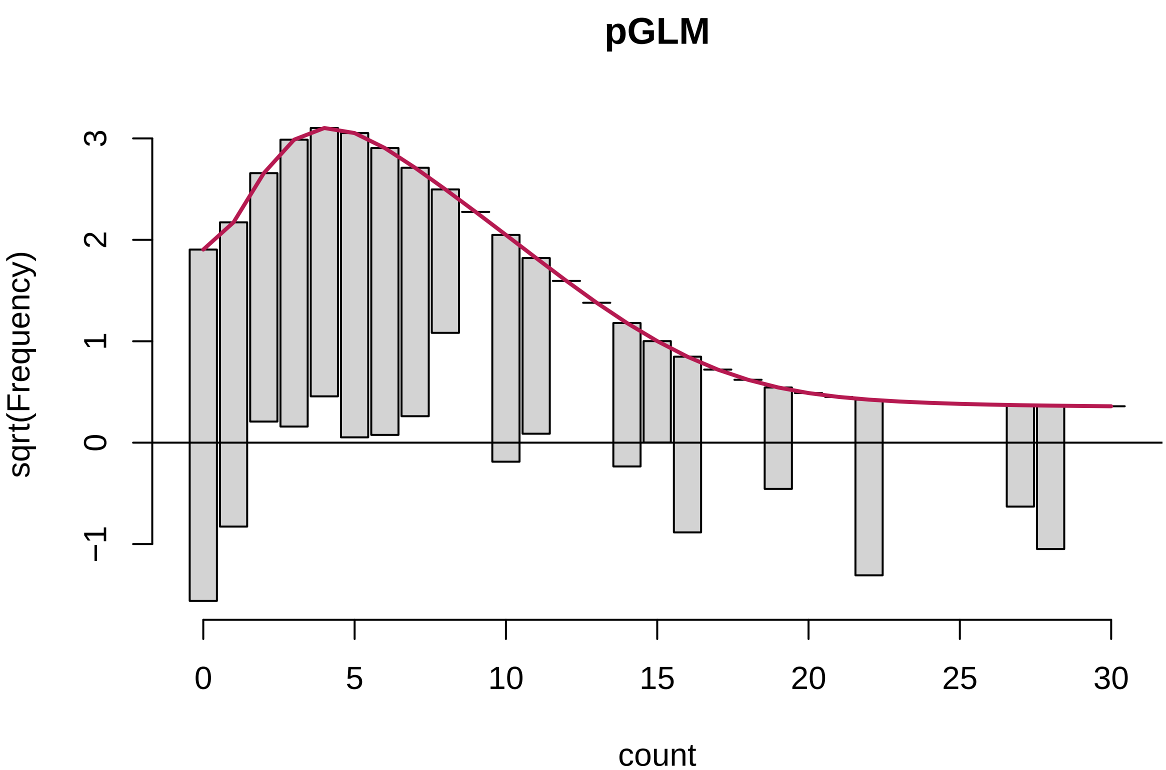
```r
# final model
pGLM <- glm(count ~ PhyloClust56 + crop_change + X30.2_PET_Mean_mm +
                cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change +
                cluster_location + mamdiv + earth11_barren,
            family = poisson(link=log), data=df.train)

summary(pGLM)
```

```
##
## Call:
## glm(formula = count ~ PhyloClust56 + crop_change + X30.2_PET_Mean_mm +
##     cluster_weather_DBSCAN + log_livestock_mam + X27.4_HuPopDen_Change +
##     cluster_location + mamdiv + earth11_barren, family = poisson(link = log),
##     data = df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9838  -1.6377  -0.5190   0.6472   4.4476
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.880e+00  1.385e+00   2.803 0.005070 **
## PhyloClust56PC3            2.176e-01  3.600e-01   0.604 0.545567
## PhyloClust56PC4           -3.355e+00  1.012e+00  -3.316 0.000914 ***
## PhyloClust56PC5           -5.079e-01  1.479e-01  -3.435 0.000593 ***
## PhyloClust56PC6            2.729e+00  4.059e-01   6.723 1.78e-11 ***
```

```
## PhyloClust56PC7            -2.911e-01  1.570e-01  -1.854 0.063733 .
## crop_change                -3.543e+01  1.341e+01  -2.643 0.008227 **
## X30.2_PET_Mean_mm           3.278e-03  4.008e-04   8.178 2.89e-16 ***
## cluster_weather_DBSCAN0     6.128e-01  3.092e-01   1.982 0.047484 *
## cluster_weather_DBSCAN1     1.296e+00  3.280e-01   3.953 7.71e-05 ***
## cluster_weather_DBSCAN2     1.500e+00  3.607e-01   4.160 3.19e-05 ***
## log_livestock_mam          -2.955e-01  9.927e-02  -2.976 0.002917 **
## X27.4_HuPopDen_Change      -6.106e+00  2.765e+00  -2.208 0.027236 *
## cluster_locationAmerica    -9.248e-01  3.301e-01  -2.801 0.005087 **
## cluster_locationAsia       -1.132e+00  2.203e-01  -5.139 2.76e-07 ***
## cluster_locationAustralia -1.866e+00  4.084e-01  -4.568 4.92e-06 ***
## cluster_locationEurope     -5.355e-01  3.503e-01  -1.529 0.126278
## mamdiv                     -2.031e-02  4.013e-03  -5.060 4.19e-07 ***
## earth11_barren             -1.727e-02  5.322e-03  -3.245 0.001176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 640.49  on 89  degrees of freedom
## Residual deviance: 350.39  on 71  degrees of freedom
## AIC: 666.81
##
## Number of Fisher Scoring iterations: 5
```

```r
rootogram(pGLM, max=30)
```



Prediction on unseen species

```r
intercept = rep(1, nrow(df.test))
Phylo_3 <- ifelse(df.test$PhyloClust56 == "PC3", 1, 0)
```

```
Phylo_4 <- ifelse(df.test$PhyloClust56 == "PC4", 1, 0)
Phylo_5 <- ifelse(df.test$PhyloClust56 == "PC5", 1, 0)
Phylo_6 <- ifelse(df.test$PhyloClust56 == "PC6", 1, 0)
Phylo_7 <- ifelse(df.test$PhyloClust56 == "PC7", 1, 0)
crop_change <- df.test$crop_change
pet <- df.test$X30.2_PET_Mean_mm
cluster_weather_0 <- ifelse(df.test$cluster_weather_DBSCAN == 0, 1, 0)
cluster_weather_1 <- ifelse(df.test$cluster_weather_DBSCAN == 1, 1, 0)
cluster_weather_2 <- ifelse(df.test$cluster_weather_DBSCAN == 2, 1, 0)
log_livestock_mam <- df.test$log_livestock_mam
HuPopChange <- df.test$X27.4_HuPopDen_Change
cluster_location_1 <- ifelse(df.test$cluster_location == "America", 1, 0)
cluster_location_2 <- ifelse(df.test$cluster_location == "Asia", 1, 0)
cluster_location_3 <- ifelse(df.test$cluster_location == "Australia", 1, 0)
cluster_location_4 <- ifelse(df.test$cluster_location == "Europe", 1, 0)
mamdiv <- df.test$mamdiv
earth11 <- df.test$earth11_barren

cm <- cbind(intercept, Phylo_3, Phylo_4, Phylo_5, Phylo_6, Phylo_7, crop_change, pet,
            cluster_weather_0, cluster_weather_1, cluster_weather_2, log_livestock_mam,
            HuPopChange, cluster_location_1, cluster_location_2, cluster_location_3, cluster_location_4
            mamdiv, earth11)


combo <- mcprofile(object=pGLM, CM=cm)

ci.result <- exp(confint(combo, level=0.95, adjust = "none"))
df.result <- data.frame(estimate=ci.result$estimate, ci = ci.result$confint)
write.csv(df.result, "ci_result.csv")
```

Build GBM for high vs low prevalence

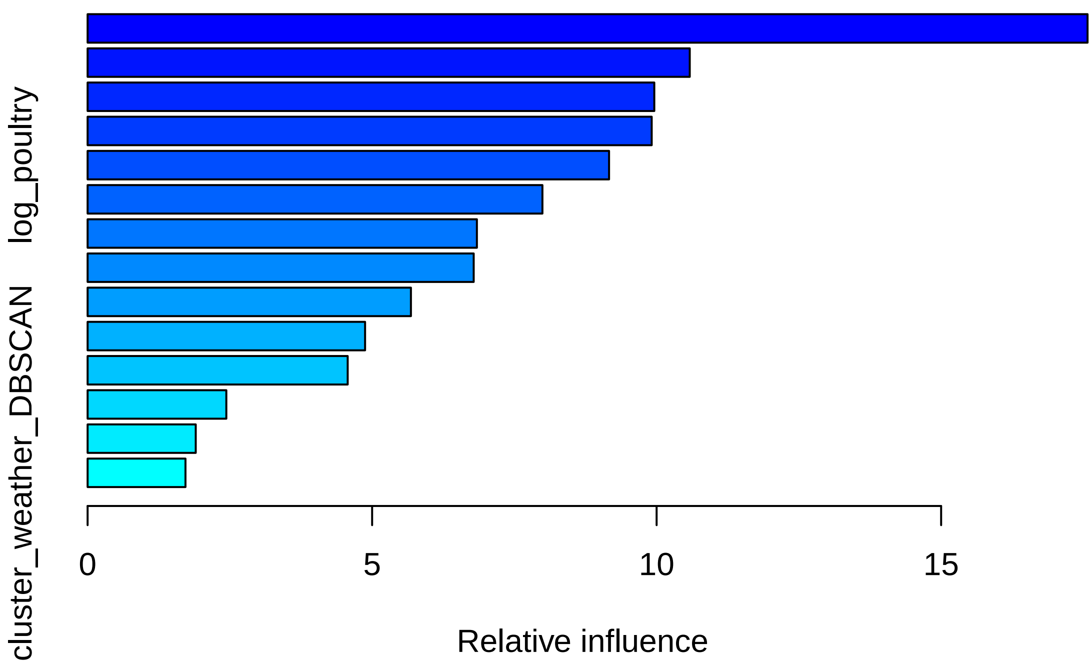
```
features = c("X27.4_HuPopDen_Change", "cluster_weather_DBSCAN", "cluster_location", "X30.1_AET_Mean_mm"
             "X30.2_PET_Mean_mm",
             "AET_divided_by_PET", "earth2_trees_everg", "crop_change", "mamdiv", "earth11_barren",
             "log_poultry", "log_livestock_mam", "earth7_veg_manag", "PhyloClust56")

GBM_model_bernoulli <- gbm(formula = label ~ . , distribution = "bernoulli",
                           data = df.train[,c("label", features)], n.trees = 50, shrinkage = 0.1,
                           interaction.depth = 4, cv.folds = 10)

print(GBM_model_bernoulli)
```

```
## gbm(formula = label ~ ., distribution = "bernoulli", data = df.train[,
##     c("label", features)], n.trees = 50, interaction.depth = 4,
##     shrinkage = 0.1, cv.folds = 10)
## A gradient boosted model with bernoulli loss function.
## 50 iterations were performed.
## The best cross-validation iteration was 11.
## There were 14 predictors of which 14 had non-zero influence.
```

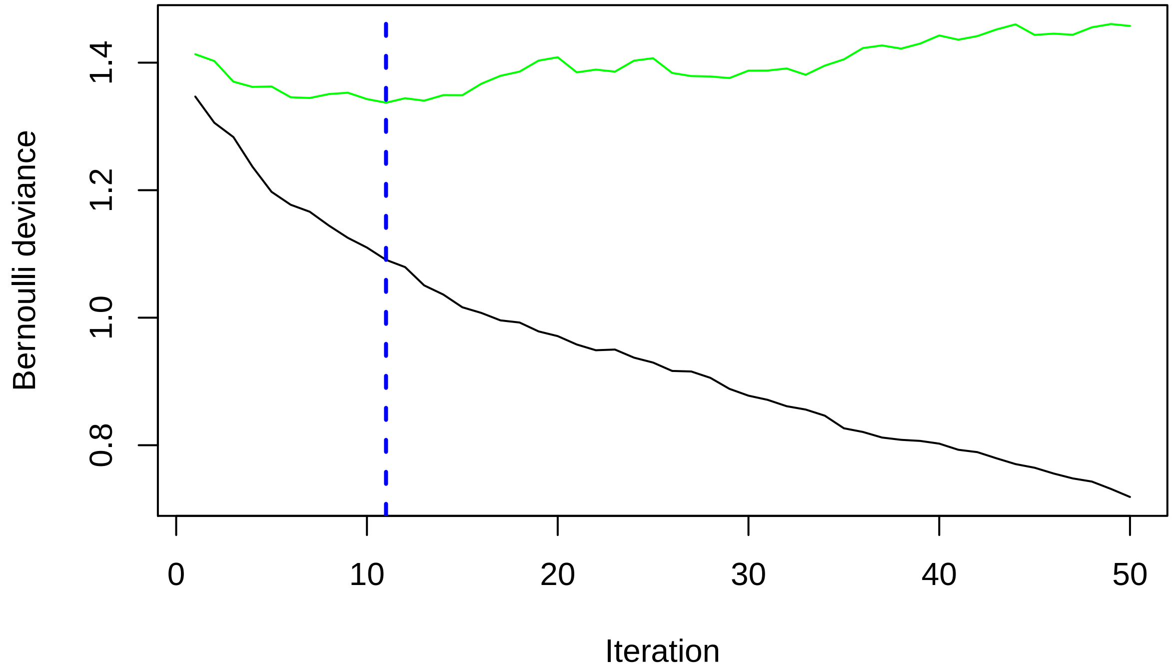
```
summary(GBM_model_bernoulli)
```




```
##                                    var   rel.inf
## mamdiv                          mamdiv 17.573444
## log_livestock_mam    log_livestock_mam 10.582340
## earth2_trees_everg  earth2_trees_everg  9.959179
## earth7_veg_manag      earth7_veg_manag  9.913460
## log_poultry                log_poultry  9.164129
## crop_change                crop_change  7.992842
## X30.2_PET_Mean_mm    X30.2_PET_Mean_mm  6.841575
## PhyloClust56              PhyloClust56  6.785507
## AET_divided_by_PET  AET_divided_by_PET  5.681845
## earth11_barren          earth11_barren  4.876376
```

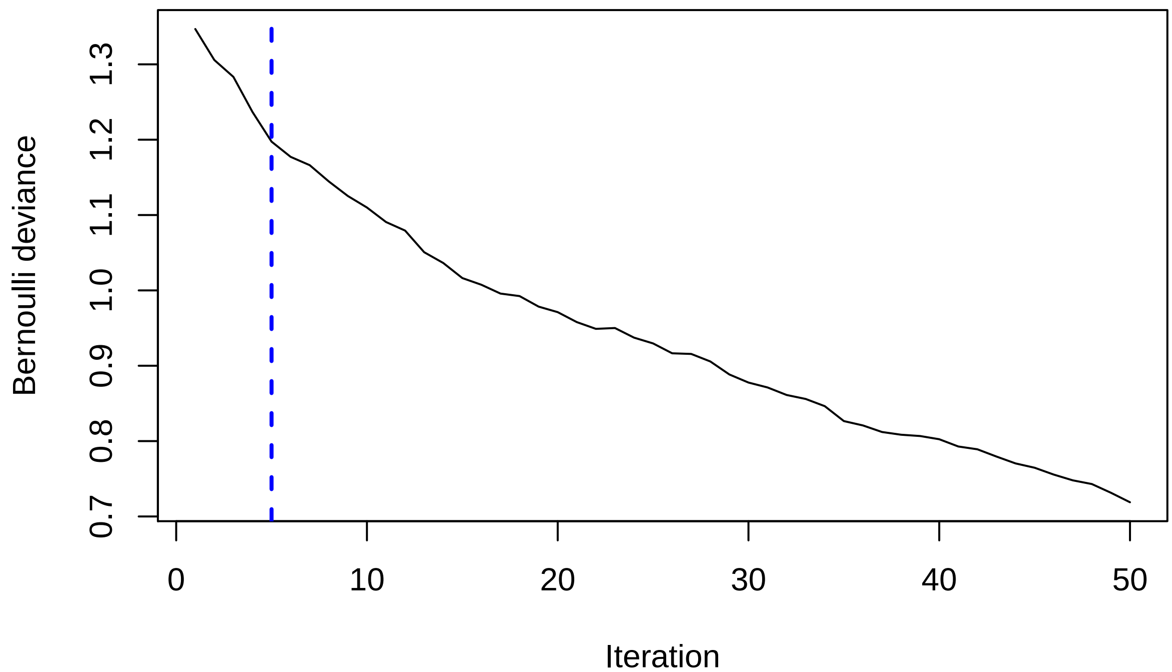
```
## X30.1_AET_Mean_mm           X30.1_AET_Mean_mm  4.570707
## cluster_location             cluster_location  2.438115
## X27.4_HuPopDen_Change   X27.4_HuPopDen_Change  1.900390
## cluster_weather_DBSCAN cluster_weather_DBSCAN  1.720090
```

```r
# plot loss function as a result of n trees added to the ensemble
optimal_cv_bernoulli <- gbm.perf(GBM_model_bernoulli, method = "cv")
```




```r
# can also test out of bag estimator
optimal_oob <- gbm.perf(GBM_model_bernoulli, method = "OOB")
```

```
## OOB generally underestimates the optimal number of iterations although predictive performance is reas
```



```r
print(optimal_cv_bernoulli)
```

```
## [1] 11
```

```r
print(optimal_oob)
```

```
## [1] 5
## attr(,"smoother")
## Call:
## loess(formula = object$oobag.improve ~ x, enp.target = min(max(4,
##     length(x)/10), 50))
##
## Number of Observations: 50
## Equivalent Number of Parameters: 4.48
## Residual Standard Error: 0.006074
```

```r
# in sample fit quality
in_sample_fit <-predict(object = GBM_model_bernoulli,
                        newdata = df.train,
                        n.trees = optimal_cv_bernoulli,
                        type = "response")
output_bernoulli <- as.factor(ifelse(in_sample_fit>0.5, 1,0))
#Train_data$CoVStatus <- as.factor(Train_data$CoVStatus)
confusionMatrix(output_bernoulli, as.factor(df.train$label))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 40 15
##          1  8 27
##
##                Accuracy : 0.7444
##                  95% CI : (0.6416, 0.8306)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 3.141e-05
##
##                   Kappa : 0.4812
##
##  Mcnemar's Test P-Value : 0.2109
##
##             Sensitivity : 0.8333
##             Specificity : 0.6429
##          Pos Pred Value : 0.7273
##          Neg Pred Value : 0.7714
##              Prevalence : 0.5333
##          Detection Rate : 0.4444
##    Detection Prevalence : 0.6111
##       Balanced Accuracy : 0.7381
##
##        'Positive' Class : 0
##
```

```r
# out of smaple fit
out_sample_fit <- predict(object = GBM_model_bernoulli,
                          newdata = df.test,
```

```
                          n.trees = optimal_cv_bernoulli,
                          type = "response")
df.GBM.result <- data.frame(binary_prediction=out_sample_fit)
write.csv(df.GBM.result, "/Volumes/D/MIDS/w210/GBM.csv")
```