Linear Models: Classification

Isabelle Kirby, Bridgette Bryant

09/16/2021

Logistic Regression:

Despite its name, in logistic regression we are classifying data. This means that the target variable is qualitative and we're trying to discern what class an observation is in. As such, linear models for logistic regression create decision boundaries to split observations into regions populated by mostly one classification. This is computationally inexpensive and works well when data can be separated cleanly (linearly). It also displays the probabilistic output in an manageable manner. This being said, a model is only as good as the data that is presented to it. When given a data set that is too small or unbalanced, the model itself isn't able to be trained well enough to be reliable for use in professional settings.It also tends to under fit data as it's not complex enough to capture non-linear decision boundaries. . . . .

Loading in file and Libraries

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ROCR)
library(mltools)
df <- read.csv("covtype.csv")
df$Cover_Type <- factor(df$Cover_Type)
```

Now we are creating the training and testing sets (80% train, 20% test).

```
i <- sample(1:nrow(df), nrow(df)*.80, replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

Data Exploration:

```
head(train)
```

```
##        Elevation Aspect Slope Horizontal_Distance_To_Hydrology
## 524748      3416    184     4                              774
## 205149      3182    354     7                              258
## 456441      3226    313    20                              108
## 464421      2830      6     2                              216
## 276307      2900    356     6                               67
## 412307      3251    191     5                              485
##        Vertical_Distance_To_Hydrology Horizontal_Distance_To_Roadways
## 524748                            -76                             810
## 205149                            -42                            2563
## 456441                             28                            2822
## 464421                              6                             741
## 276307                             -2                             942
## 412307                            120                            1689
##        Hillshade_9am Hillshade_Noon Hillshade_3pm
## 524748           220            242           157
## 205149           209            228           157
## 456441           162            220           196
## 464421           216            234           155
## 276307           210            229           157
## 412307           220            244           160
##        Horizontal_Distance_To_Fire_Points Wilderness_Area1 Wilderness_Area2
## 524748                                362                0                1
## 205149                               2293                1                0
## 456441                               1266                0                0
## 464421                                552                0                0
## 276307                               2831                1                0
## 412307                               2301                0                0
##        Wilderness_Area3 Wilderness_Area4 Soil_Type Soil_Type1 Soil_Type2
## 524748                0                0        40          0          0
## 205149                0                0        22          0          0
## 456441                1                0        23          0          0
## 464421                1                0        23          0          0
## 276307                0                0        22          0          0
## 412307                1                0        31          0          0
##        Soil_Type3 Soil_Type4 Soil_Type5 Soil_Type6 Soil_Type7 Soil_Type8
## 524748          0          0          0          0          0          0
## 205149          0          0          0          0          0          0
## 456441          0          0          0          0          0          0
## 464421          0          0          0          0          0          0
## 276307          0          0          0          0          0          0
## 412307          0          0          0          0          0          0
##        Soil_Type9 Soil_Type10 Soil_Type11 Soil_Type12 Soil_Type13 Soil_Type14
## 524748          0           0           0           0           0           0
## 205149          0           0           0           0           0           0
## 456441          0           0           0           0           0           0
## 464421          0           0           0           0           0           0
## 276307          0           0           0           0           0           0
## 412307          0           0           0           0           0           0
##        Soil_Type15 Soil_Type16 Soil_Type17 Soil_Type18 Soil_Type19 Soil_Type20
## 524748           0           0           0           0           0           0
## 205149           0           0           0           0           0           0
## 456441           0           0           0           0           0           0
## 464421           0           0           0           0           0           0
## 276307           0           0           0           0           0           0
## 412307           0           0           0           0           0           0
##        Soil_Type21 Soil_Type22 Soil_Type23 Soil_Type24 Soil_Type25 Soil_Type26
## 524748           0           0           0           0           0           0
## 205149           0           1           0           0           0           0
## 456441           0           0           1           0           0           0
## 464421           0           0           1           0           0           0
## 276307           0           1           0           0           0           0
## 412307           0           0           0           0           0           0
##        Soil_Type27 Soil_Type28 Soil_Type29 Soil_Type30 Soil_Type31 Soil_Type32
## 524748           0           0           0           0           0           0
## 205149           0           0           0           0           0           0
## 456441           0           0           0           0           0           0
## 464421           0           0           0           0           0           0
## 276307           0           0           0           0           0           0
## 412307           0           0           0           0           1           0
##        Soil_Type33 Soil_Type34 Soil_Type35 Soil_Type36 Soil_Type37 Soil_Type38
## 524748           0           0           0           0           0           0
## 205149           0           0           0           0           0           0
## 456441           0           0           0           0           0           0
## 464421           0           0           0           0           0           0
## 276307           0           0           0           0           0           0
## 412307           0           0           0           0           0           0
##        Soil_Type39 Soil_Type40 Cover_Type
## 524748           0           1          7
## 205149           0           0          1
## 456441           0           0          1
## 464421           0           0          1
## 276307           0           0          1
## 412307           0           0          1
```

```
tail(train)
```

```
##        Elevation Aspect Slope Horizontal_Distance_To_Hydrology
## 153760      3120    163    10                              218
## 168874      2794      7    16                              297
## 87326       3047    247     7                              811
## 393316      2951    138    14                              741
## 203005      3266    318     6                              180
## 166072      2689     17    13                               42
##        Vertical_Distance_To_Hydrology Horizontal_Distance_To_Roadways
## 153760                             24                            5418
## 168874                             55                             474
## 87326                              48                            5788
## 393316                             75                            2839
## 203005                             23                            5583
## 166072                              1                             162
##        Hillshade_9am Hillshade_Noon Hillshade_3pm
## 153760           230            243           145
## 168874           199            207           143
## 87326            205            246           179
## 393316           240            235           121
## 203005           205            235           169
## 166072           208            212           139
##        Horizontal_Distance_To_Fire_Points Wilderness_Area1 Wilderness_Area2
## 153760                                808                1                0
## 168874                               2352                1                0
## 87326                                3611                1                0
## 393316                               1767                0                0
## 203005                                484                1                0
## 166072                               2663                1                0
##        Wilderness_Area3 Wilderness_Area4 Soil_Type Soil_Type1 Soil_Type2
## 153760                0                0        29          0          0
## 168874                0                0        24          0          0
## 87326                 0                0        29          0          0
## 393316                1                0        26          0          0
## 203005                0                0        38          0          0
## 166072                0                0        24          0          0
##        Soil_Type3 Soil_Type4 Soil_Type5 Soil_Type6 Soil_Type7 Soil_Type8
## 153760          0          0          0          0          0          0
## 168874          0          0          0          0          0          0
## 87326           0          0          0          0          0          0
## 393316          0          0          0          0          0          0
## 203005          0          0          0          0          0          0
## 166072          0          0          0          0          0          0
##        Soil_Type9 Soil_Type10 Soil_Type11 Soil_Type12 Soil_Type13 Soil_Type14
## 153760          0           0           0           0           0           0
## 168874          0           0           0           0           0           0
## 87326           0           0           0           0           0           0
## 393316          0           0           0           0           0           0
## 203005          0           0           0           0           0           0
## 166072          0           0           0           0           0           0
##        Soil_Type15 Soil_Type16 Soil_Type17 Soil_Type18 Soil_Type19 Soil_Type20
## 153760           0           0           0           0           0           0
## 168874           0           0           0           0           0           0
## 87326            0           0           0           0           0           0
## 393316           0           0           0           0           0           0
## 203005           0           0           0           0           0           0
## 166072           0           0           0           0           0           0
##        Soil_Type21 Soil_Type22 Soil_Type23 Soil_Type24 Soil_Type25 Soil_Type26
## 153760           0           0           0           0           0           0
## 168874           0           0           0           1           0           0
## 87326            0           0           0           0           0           0
## 393316           0           0           0           0           0           1
## 203005           0           0           0           0           0           0
## 166072           0           0           0           1           0           0
##        Soil_Type27 Soil_Type28 Soil_Type29 Soil_Type30 Soil_Type31 Soil_Type32
## 153760           0           0           1           0           0           0
## 168874           0           0           0           0           0           0
## 87326            0           0           1           0           0           0
## 393316           0           0           0           0           0           0
## 203005           0           0           0           0           0           0
## 166072           0           0           0           0           0           0
##        Soil_Type33 Soil_Type34 Soil_Type35 Soil_Type36 Soil_Type37 Soil_Type38
## 153760           0           0           0           0           0           0
## 168874           0           0           0           0           0           0
## 87326            0           0           0           0           0           0
## 393316           0           0           0           0           0           0
## 203005           0           0           0           0           0           1
## 166072           0           0           0           0           0           0
##        Soil_Type39 Soil_Type40 Cover_Type
## 153760           0           0          1
## 168874           0           0          2
## 87326            0           0          1
## 393316           0           0          2
## 203005           0           0          2
## 166072           0           0          2
```

```
median(train$Elevation)
```

```
## [1] 2996
```

```r
getmode <- function(d) {
  uniqd <- unique(d)
  uniqd[which.max(tabulate(match(d, uniqd)))]
}
getmode(train$Cover_Type)
```

```
## [1] 2
## Levels: 1 2 3 4 5 6 7
```

```r
getmode(train$Soil_Type)
```

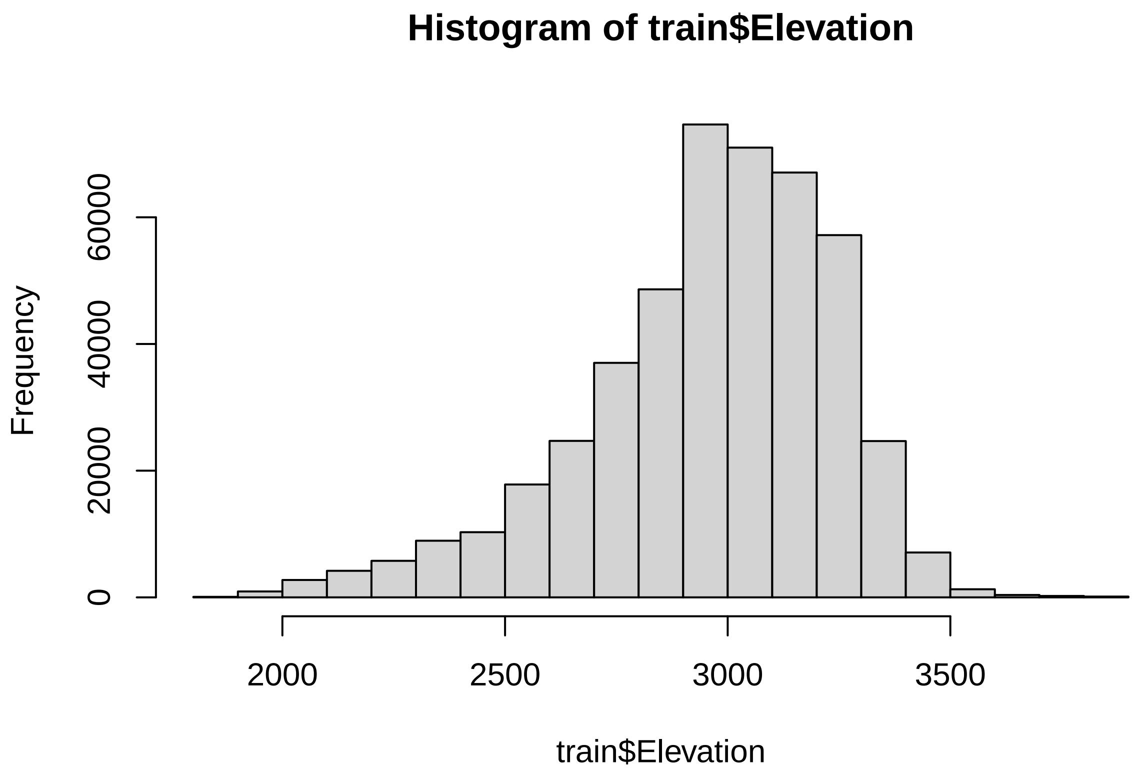
```
## [1] 29
```

```r
str(train)
```

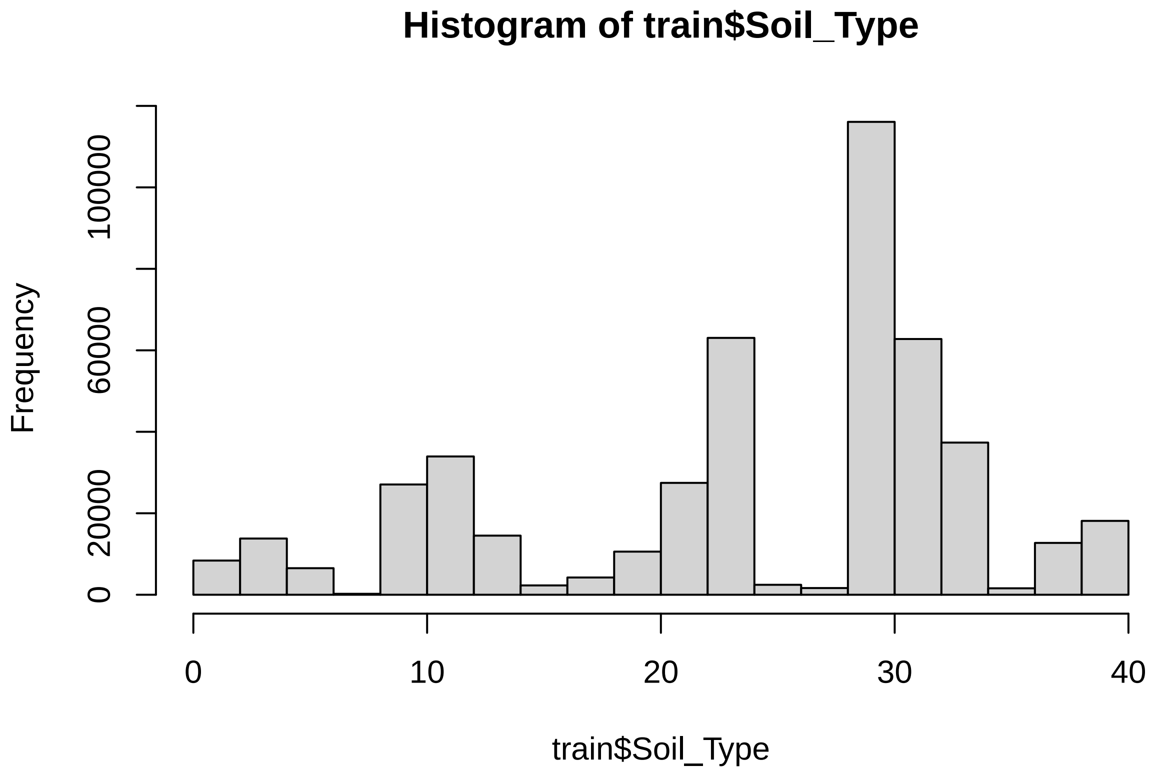
```
## 'data.frame':    464809 obs. of  56 variables:
##  $ Elevation                         : int  3416 3182 3226 2830 2900 3251 2772 2847 3000 3289 ...
##  $ Aspect                            : int  184 354 313 6 356 191 158 223 33 24 ...
##  $ Slope                             : int  4 7 20 2 6 5 20 21 8 6 ...
##  $ Horizontal_Distance_To_Hydrology  : int  774 258 108 216 67 485 42 85 216 0 ...
##  $ Vertical_Distance_To_Hydrology    : int  -76 -42 28 6 -2 120 10 22 -13 0 ...
##  $ Horizontal_Distance_To_Roadways   : int  810 2563 2822 741 942 1689 4002 2306 3318 2125 ...
##  $ Hillshade_9am                     : int  220 209 162 216 210 220 237 188 219 217 ...
##  $ Hillshade_Noon                    : int  242 228 220 234 229 244 239 254 223 227 ...
##  $ Hillshade_3pm                     : int  157 157 196 155 157 160 122 196 140 147 ...
##  $ Horizontal_Distance_To_Fire_Points: int  362 2293 1266 552 2831 2301 5460 2012 5511 2067 ...
##  $ Wilderness_Area1                  : int  0 1 0 0 1 0 1 0 1 0 ...
##  $ Wilderness_Area2                  : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ Wilderness_Area3                  : int  0 0 1 1 0 1 0 1 0 0 ...
##  $ Wilderness_Area4                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type                         : int  40 22 23 23 22 31 30 13 29 38 ...
##  $ Soil_Type1                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type2                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type3                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type4                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type5                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type6                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type7                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type8                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type9                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type10                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type11                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type12                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type13                       : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Soil_Type14                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type15                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type16                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type17                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type18                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type19                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type20                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type21                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type22                       : int  0 1 0 0 1 0 0 0 0 0 ...
##  $ Soil_Type23                       : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ Soil_Type24                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type25                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type26                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type27                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type28                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type29                       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Soil_Type30                       : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Soil_Type31                       : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Soil_Type32                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type33                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type34                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type35                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type36                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type37                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type38                       : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Soil_Type39                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soil_Type40                       : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Cover_Type                        : Factor w/ 7 levels "1","2","3","4",..: 7 1 1 1 1 1 2 2 2 2 ..
```

Plots

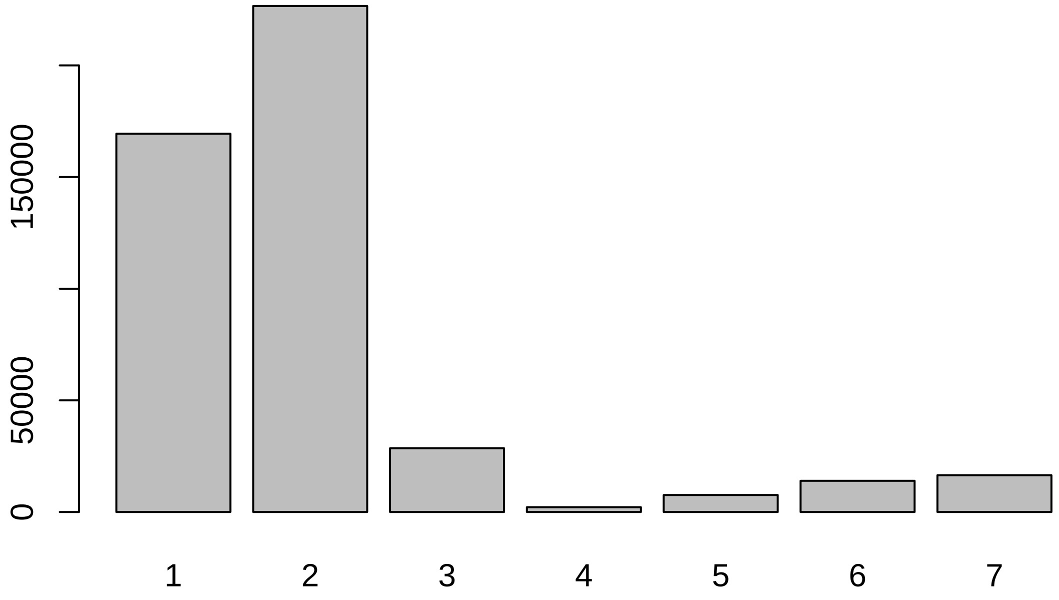
```
hist(train$Elevation)
```



```
hist(train$Soil_Type)
```

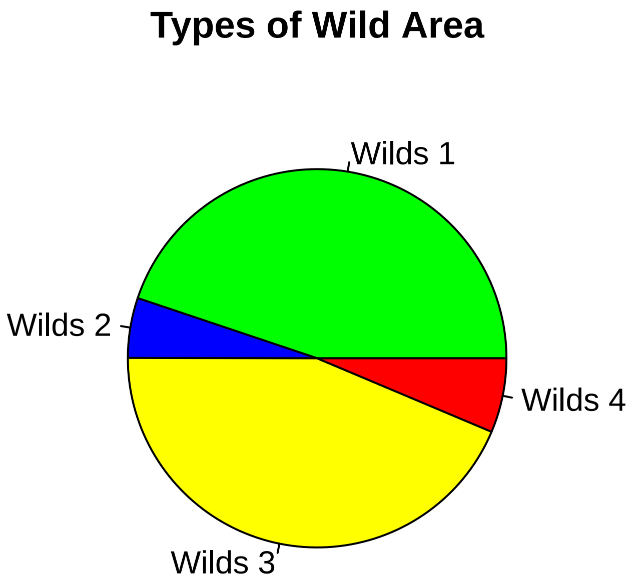


```
counts <- table(train$Cover_Type)
barplot(counts)
```



```
slices <- c(sum(train$Wilderness_Area1==1, na.rm=TRUE), sum(train$Wilderness_Area2==1, na.rm=TRUE), sum

labls <- c("Wilds 1","Wilds 2","Wilds 3","Wilds 4")
pie(slices,labels = labls, main="Types of Wild Area", col=c("green", "blue", "yellow", "red"))
```



## Creating the One vs All Models:

```
tree_type1 <- df
tree_type1$Cover_Type <- as.factor(ifelse (tree_type1$Cover_Type=="1",1,0))

tree_type2 <- df
tree_type2$Cover_Type <- as.factor(ifelse (tree_type2$Cover_Type=="2",1,0))

tree_type3 <- df
tree_type3$Cover_Type <- as.factor(ifelse (tree_type3$Cover_Type=="3",1,0))

tree_type4 <- df
tree_type4$Cover_Type <- as.factor(ifelse (tree_type4$Cover_Type=="4",1,0))

tree_type5 <- df
tree_type5$Cover_Type <- as.factor(ifelse (tree_type5$Cover_Type=="5",1,0))

tree_type6 <- df
tree_type6$Cover_Type <- as.factor(ifelse (tree_type6$Cover_Type=="6",1,0))

tree_type7 <- df
tree_type7$Cover_Type <- as.factor(ifelse (tree_type7$Cover_Type=="7",1,0))

fun <- function(dataf, i){
  funtrain <- dataf[i,]
  funtest <- dataf[-i,]
  glm1 <- glm(Cover_Type~., data=funtrain, family="binomial")
  probs <- (predict(glm1, newdata=funtest))
  pred <- ifelse(probs>0.5, 1, 0)
  confmatrix <- confusionMatrix(data=as.factor(pred), reference=funtest$Cover_Type)
  print(confmatrix)

  p <- predict(glm1, newdata=funtest, type="response")
  pr <- prediction(p, funtest$Cover_Type)
  prf <- performance(pr, measure="tpr", x.measure="fpr")
  plot(prf)

  auc <- performance(pr, measure="auc")
  auc <- auc@y.values[[1]]
  print(paste("AUC: ", auc))

  #mcc(confusionM= as.matrix(confmatrix))

}

fun(tree_type1, i)
```

```
## Warning: glm.fit: algorithm did not converge
```

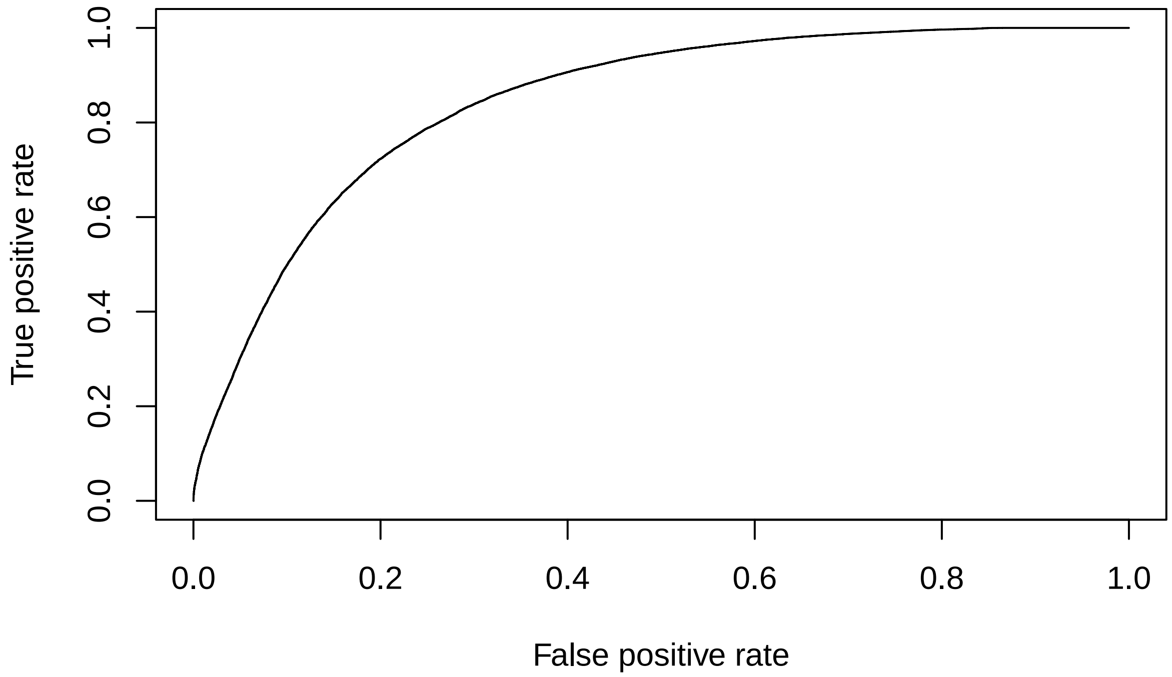
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 67034 22600
##          1  6718 19851
##
##                Accuracy : 0.7477
##                  95% CI : (0.7452, 0.7502)
##     No Information Rate : 0.6347
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.409
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9089
##             Specificity : 0.4676
##          Pos Pred Value : 0.7479
##          Neg Pred Value : 0.7471
##              Prevalence : 0.6347
##          Detection Rate : 0.5769
##    Detection Prevalence : 0.7714
##       Balanced Accuracy : 0.6883
##
##        'Positive' Class : 0
##
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
```

```
## prediction from a rank-deficient fit may be misleading
```



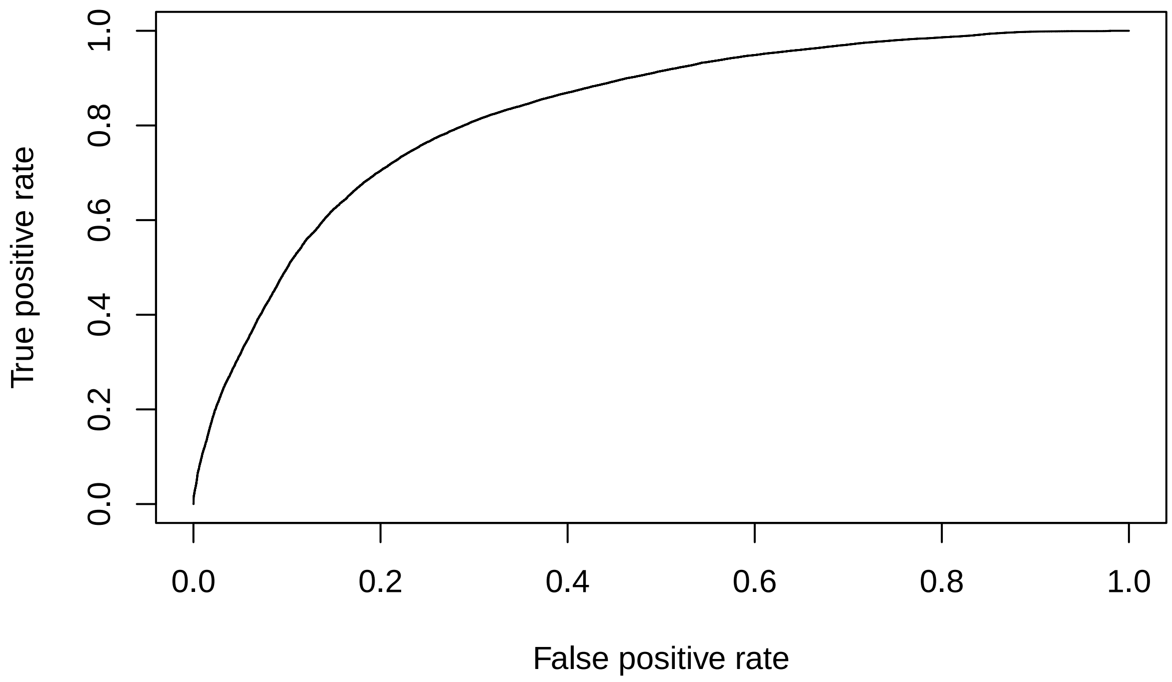
```
## [1] "AUC:  0.843731028052134"
```

```
fun(tree_type2, i)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 51143 22484
##          1  8375 34201
##
##                Accuracy : 0.7344
##                  95% CI : (0.7319, 0.737)
##     No Information Rate : 0.5122
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4654
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8593
##             Specificity : 0.6034
##          Pos Pred Value : 0.6946
##          Neg Pred Value : 0.8033
##              Prevalence : 0.5122
##          Detection Rate : 0.4401
##    Detection Prevalence : 0.6336
##       Balanced Accuracy : 0.7313
##
##        'Positive' Class : 0
##
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```



```
## [1] "AUC:  0.827823830948537"
```

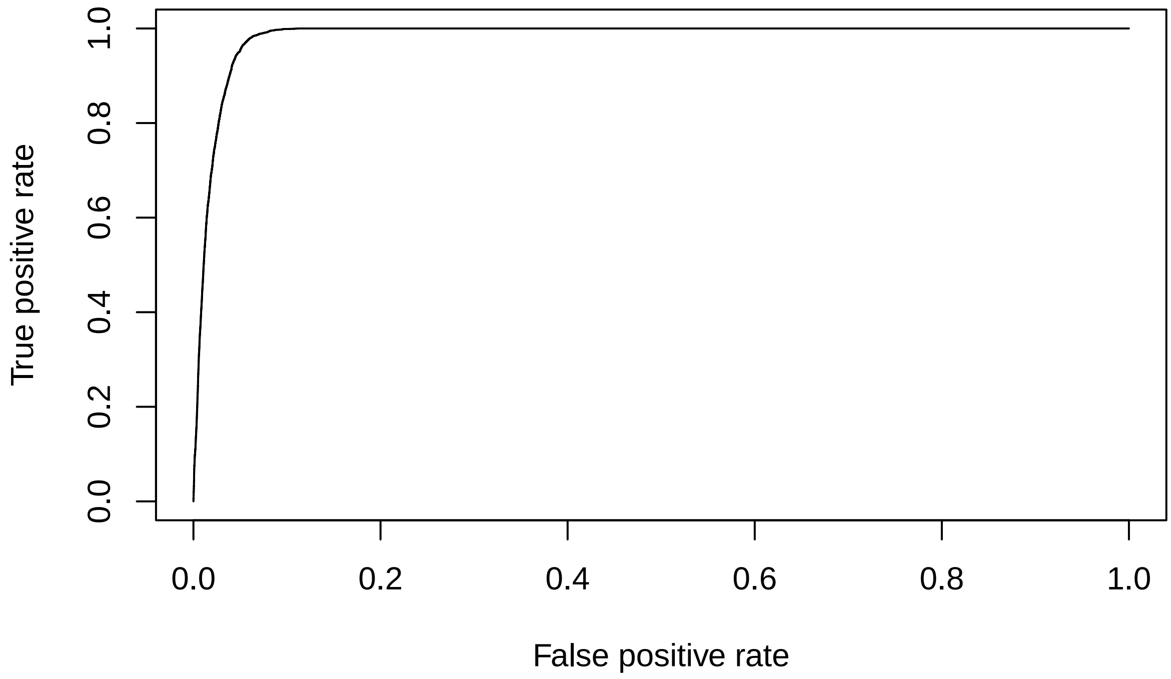
```
fun(tree_type3, i)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: prediction from a rank-deficient fit may be misleading
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 107796   3566
##          1   1215   3626
##
##                Accuracy : 0.9589
##                  95% CI : (0.9577, 0.96)
##     No Information Rate : 0.9381
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5819
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9889
##             Specificity : 0.5042
##          Pos Pred Value : 0.9680
##          Neg Pred Value : 0.7490
##              Prevalence : 0.9381
##          Detection Rate : 0.9277
##    Detection Prevalence : 0.9583
##       Balanced Accuracy : 0.7465
##
##        'Positive' Class : 0
##
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```



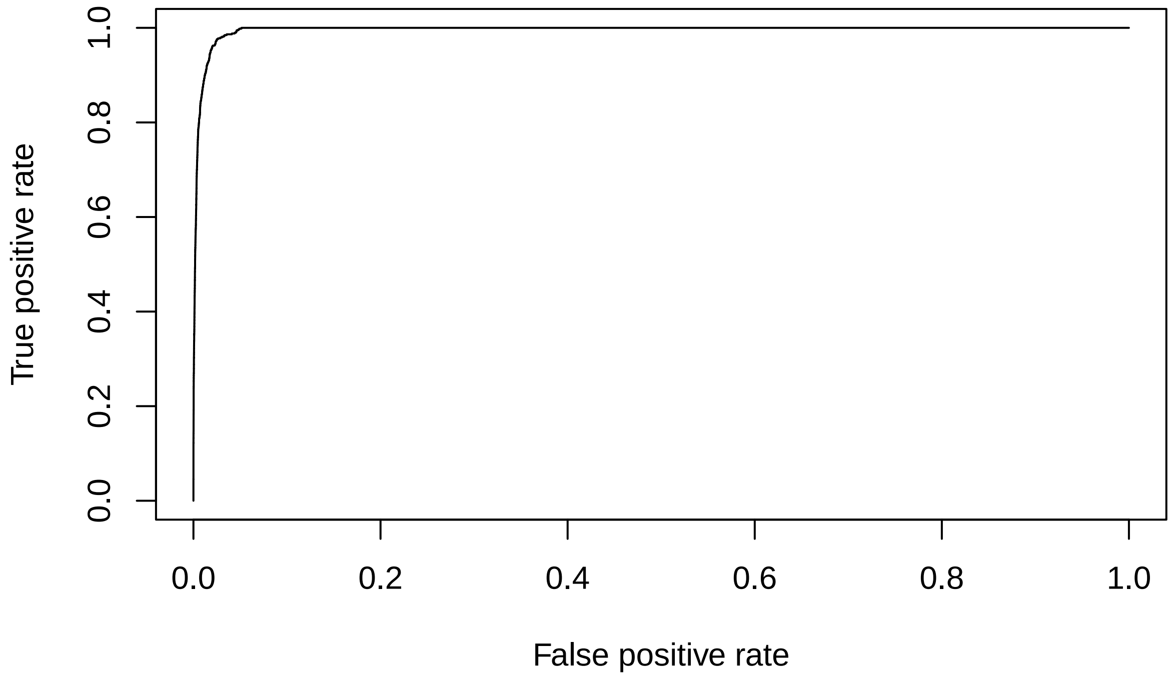
```
## [1] "AUC:  0.983775061724187"
```

```
fun(tree_type4, i)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: prediction from a rank-deficient fit may be misleading
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 115576    431
##          1     41    155
##
##                Accuracy : 0.9959
##                  95% CI : (0.9956, 0.9963)
##     No Information Rate : 0.995
##     P-Value [Acc > NIR] : 5.926e-07
##
##                   Kappa : 0.3949
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9996
##             Specificity : 0.2645
##          Pos Pred Value : 0.9963
##          Neg Pred Value : 0.7908
##              Prevalence : 0.9950
##          Detection Rate : 0.9946
##    Detection Prevalence : 0.9983
##       Balanced Accuracy : 0.6321
##
##        'Positive' Class : 0
##
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```



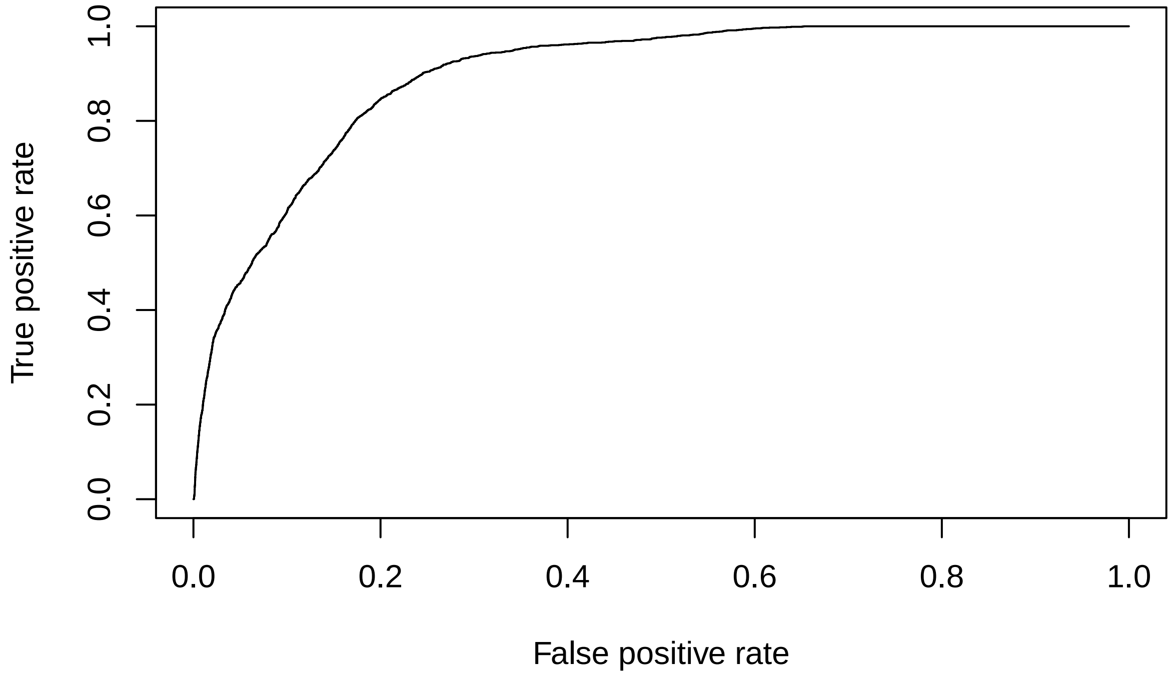
```
## [1] "AUC:  0.99557139361599"
```

```
fun(tree_type5, i)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 114289   1875
##          1     39      0
##
##                Accuracy : 0.9835
##                  95% CI : (0.9828, 0.9843)
##     No Information Rate : 0.9839
##     P-Value [Acc > NIR] : 0.8213
##
##                   Kappa : -7e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9997
##             Specificity : 0.0000
##          Pos Pred Value : 0.9839
##          Neg Pred Value : 0.0000
##               Prevalence : 0.9839
##           Detection Rate : 0.9835
##     Detection Prevalence : 0.9997
##        Balanced Accuracy : 0.4998
##
##         'Positive' Class : 0
##
```



```
## [1] "AUC:  0.896389587852439"
```

```
fun(tree_type6, i)
```

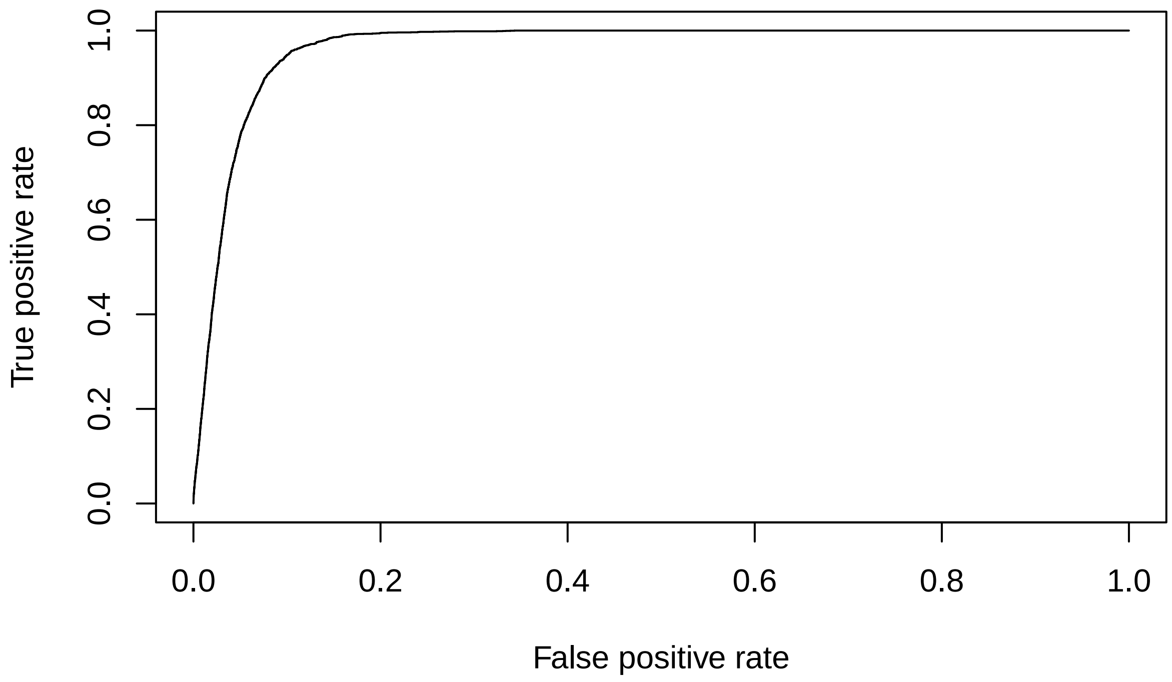
```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 112727   3271
##          1     97    108
##
##                Accuracy : 0.971
##                  95% CI : (0.97, 0.972)
##      No Information Rate : 0.9709
##      P-Value [Acc > NIR] : 0.4283
##
##                    Kappa : 0.0571
##
##   Mcnemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.99914
##              Specificity : 0.03196
##           Pos Pred Value : 0.97180
##           Neg Pred Value : 0.52683
##               Prevalence : 0.97092
##           Detection Rate : 0.97009
##     Detection Prevalence : 0.99824
##        Balanced Accuracy : 0.51555
##
##         'Positive' Class : 0
##
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```



```
## [1] "AUC:  0.964593947990211"
```

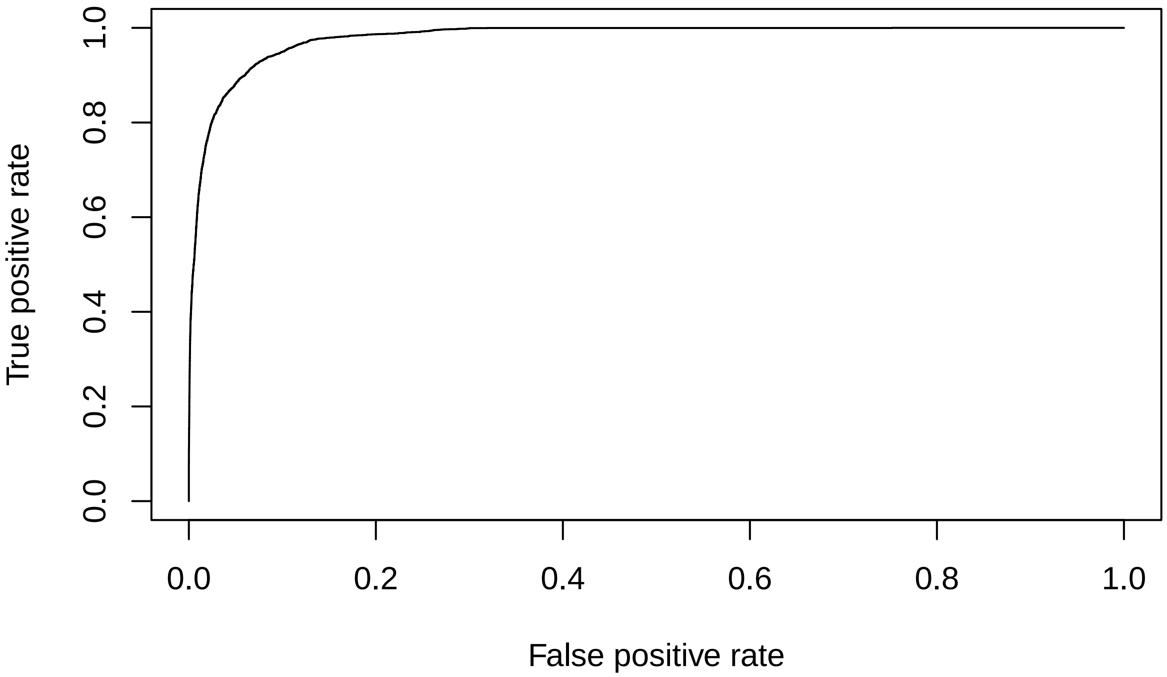
```
fun(tree_type7, i)
```

```
## Warning: glm.fit: algorithm did not converge
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 111664   2086
##          1    504   1949
##
##                Accuracy : 0.9777
##                  95% CI : (0.9768, 0.9786)
##     No Information Rate : 0.9653
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.59
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9955
##             Specificity : 0.4830
##          Pos Pred Value : 0.9817
##          Neg Pred Value : 0.7945
##              Prevalence : 0.9653
##          Detection Rate : 0.9609
##    Detection Prevalence : 0.9789
##       Balanced Accuracy : 0.7393
##
##        'Positive' Class : 0
##

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```



```
## [1] "AUC:  0.979808542187152"
```

Logistic Regression Model Summary and Evaluation:

Our logistical regression has relatively high accuracy rates for all of the different types of trees. Since there are seven different types of trees present in the data, we split the data solely about the tree type into a dataframe specifically made for them. From there we made models for each tree type.

Type 1: For this model we had about a 75% accuracy level. Most of the trees were not of type 1, and the model represented that accordingly. It did have more false positives than negatives, meaning that the growing needs of this tree have a slightly high variance. This would be true of common trees. Type 2: For this model we had an accuracy level of about 73%, meaning that our model was mostly accurate. It also had more difficulty with false positives than it did with false negatives. This again may be due to that tree having less specific needs for its growth and having a high variance in the data. There were a lot of true positives, meaning that this tree is the most common tree in the data set.

Type 3: This model had an accuracy level of about 96%, a truly impressive accuracy level. It also had a more difficult time with identifying more false positives than false negatives, but that could be due to the common trees in type 1 and type 2. This tree most likely has some specific needs (ie: a lower variance in the data). This does mean that the bias is higher in this model, however it is still incredibly accurate. This also makes sense considering how few of the trees were true positives, meaning this tree is uncommon.

Type 4: This model had an accuracy of 99.6%. It identified a lot of trees as correctly being apart of that tree type. It seems that it had more of an issue with false positives than it did false negatives, but it still accurately described the data set given. There were very few trees of this type in the data set and seemed to have very specific needs as the model had an very few false negatives and positives.

Type 5: For this model we have a 98% accuracy level. This is mostly because there were very few trees of type five in the data set, making it very hard to identify what trees were apart of type 5. With a data set as large as ours, however, the model is still mostly accurate as there are so few trees of type 5.

Type 6: For this model we have nearly a 97% accuracy level. It is also identified more true negatives than anything else, meaning that there weren't many trees of type 6. This model had very few false positives, but there were still more of them then true positives. This can be attributed to some other tree types being more common and having few specifications for their growth.

Type 7: This model had around a 98% accuracy level. This is because there are few trees of type 7 in the data set. There were more errors than there were true positives, meaning that the data for the tree wasn't specific enough to help differentiate the few trees from the majority of trees present.

On average the overall accuracy is: 90.9%

Creating the Naive Bayes Model:

```
library(e1071)
```

```
##
## Attaching package: 'e1071'
```

```
## The following object is masked from 'package:mltools':
##
##     skewness
```

```
nb1 <- naiveBayes(Cover_Type~., data=train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##           1           2           3           4           5           6
## 0.364427109 0.487546498 0.061448896 0.004649222 0.016389528 0.030094082
##           7
## 0.035444667
##
## Conditional probabilities:
##    Elevation
## Y       [,1]     [,2]
##   1 3128.507 157.6738
##   2 2921.116 186.5778
##   3 2394.754 196.9471
##   4 2222.234 102.7698
##   5 2787.953  96.4519
##   6 2420.121 188.3042
##   7 3361.445 106.9748
##
##    Aspect
## Y       [,1]      [,2]
##   1 156.3371 116.79555
##   2 151.9960 107.59632
##   3 176.6000 107.76867
##   4 138.6293  88.32668
##   5 140.0320  91.95264
##   6 180.8822 133.80690
##   7 153.2058 110.80917
##
##    Slope
## Y       [,1]     [,2]
##   1 13.12711 6.826340
##   2 13.54897 7.101948
##   3 20.78345 9.018799
##   4 18.37483 9.360634
##   5 16.59740 8.180999
##   6 19.02681 7.917900
##   7 14.23678 7.468867
##
##    Horizontal_Distance_To_Hydrology
## Y       [,1]     [,2]
##   1 270.9312 216.5833
##   2 280.2213 210.3954
##   3 210.4634 142.0389
##   4 105.5044 138.6233
##   5 211.5620 178.1530
##   6 160.3096 124.5471
##   7 357.3151 295.0068
##
##    Vertical_Distance_To_Hydrology
## Y       [,1]     [,2]
##   1 42.19253 56.60597
##   2 45.92361 57.56806
##   3 62.73636 59.04611
##   4 40.43915 58.63341
##   5 50.76700 57.90432
##   6 45.54146 46.95561
##   7 69.64437 79.99382
##
##    Horizontal_Distance_To_Roadways
## Y        [,1]      [,2]
##   1 2616.7469 1499.8758
##   2 2429.6413 1617.4292
##   3  944.3041  614.5095
##   4  916.9403  365.1336
##   5 1352.5045 1048.5228
##   6 1039.2190  570.6212
##   7 2731.7975 1199.4858
##
##    Hillshade_9am
## Y       [,1]     [,2]
##   1 211.9711 24.85436
##   2 213.8605 24.89907
##   3 201.7603 40.64581
##   4 227.8339 24.46094
##   5 223.3505 22.74593
##   6 192.7881 33.56233
##   7 216.8626 23.44931
```

```
##
##    Hillshade_Noon
## Y       [,1]     [,2]
##   1 223.4077 18.14638
##   2 225.3448 18.51479
##   3 215.8135 27.93759
##   4 217.4452 20.78331
##   5 219.1890 24.80578
##   6 209.9790 24.26590
##   7 221.7173 20.11745
##
##    Hillshade_3pm
## Y       [,1]     [,2]
##   1 143.8885 36.06475
##   2 142.9789 36.24412
##   3 140.5359 52.40727
##   4 112.5988 49.27332
##   5 122.2931 49.22269
##   6 148.5394 45.21777
##   7 135.0682 38.81726
##
##    Horizontal_Distance_To_Fire_Points
## Y        [,1]      [,2]
##   1 2009.4115 1236.1904
##   2 2166.7788 1422.2615
##   3  909.5973  526.6107
##   4  853.4419  477.8822
##   5 1569.9078  988.4897
##   6 1053.5453  577.7092
##   7 2070.0090 1088.2663
##
##    Wilderness_Area1
## Y        [,1]      [,2]
##   1 0.4995011 0.5000012
##   2 0.5152505 0.4997685
##   3 0.0000000 0.0000000
##   4 0.0000000 0.0000000
##   5 0.3951168 0.4889079
##   6 0.0000000 0.0000000
##   7 0.2468589 0.4311970
##
##    Wilderness_Area2
## Y         [,1]      [,2]
##   1 0.08740237 0.2824246
##   2 0.03161295 0.1749677
##   3 0.00000000 0.0000000
##   4 0.00000000 0.0000000
##   5 0.00000000 0.0000000
##   6 0.00000000 0.0000000
##   7 0.11180577 0.3151369
##
##    Wilderness_Area3
## Y        [,1]      [,2]
##   1 0.4130965 0.4923913
##   2 0.4424092 0.4966733
##   3 0.4015825 0.4902269
##   4 0.0000000 0.0000000
##   5 0.6048832 0.4889079
##   6 0.4396626 0.4963638
##   7 0.6413354 0.4796231
##
##    Wilderness_Area4
## Y         [,1]      [,2]
##   1 0.0000000 0.0000000
##   2 0.0107274 0.1030163
##   3 0.5984175 0.4902269
##   4 1.0000000 0.0000000
##   5 0.0000000 0.0000000
##   6 0.5603374 0.4963638
##   7 0.0000000 0.0000000
##
##    Soil_Type
## Y         [,1]     [,2]
##   1 27.756608 6.153684
##   2 24.346414 8.396812
##   3  6.297003 3.969614
##   4  7.169366 5.643089
##   5 21.620110 9.601763
##   6 10.195382 6.595261
##   7 36.599272 4.781332
##
##    Soil_Type1
## Y          [,1]      [,2]
##   1 0.00000000 0.0000000
##   2 0.00000000 0.0000000
##   3 0.05906449 0.2381139
##   4 0.06848681 0.2616401
##   5 0.00000000 0.0000000
##   6 0.04446669 0.2102577
##   7 0.00000000 0.0000000
##
##    Soil_Type2
## Y           [,1]       [,2]
##   1 0.000000000 0.00000000
##   2 0.002965369 0.05437452
##   3 0.139451019 0.34642262
##   4 0.039796391 0.19552578
##   5 0.027960095 0.16486933
##   6 0.076994567 0.26659236
##   7 0.000000000 0.00000000
##
##    Soil_Type3
## Y           [,1]       [,2]
##   1 0.000000000 0.00000000
##   2 0.004183288 0.06454306
##   3 0.068237518 0.25215746
##   4 0.365108746 0.48157208
##   5 0.000000000 0.00000000
##   6 0.011509866 0.10666866
##   7 0.000000000 0.00000000
## 
##    Soil_Type4
## Y           [,1]       [,2]
##   1 0.0008501142 0.02914441
##   2 0.0115261058 0.10673943
##   3 0.2095441496 0.40699042
##   4 0.0624710782 0.24206519
##   5 0.0630086637 0.24299449
##   6 0.0353874750 0.18476375
##   7 0.0039453718 0.06269007
## 
##    Soil_Type5
## Y         [,1]      [,2]
##   1 0.00000000 0.0000000
##   2 0.00000000 0.0000000
##   3 0.02706393 0.1622726
##   4 0.01573346 0.1244712
##   5 0.00000000 0.0000000
##   6 0.03302831 0.1787169
##   7 0.00000000 0.0000000
## 
##    Soil_Type6
## Y          [,1]       [,2]
##   1 0.000000000 0.00000000
##   2 0.003230134 0.05674253
##   3 0.111511799 0.31477037
##   4 0.114298936 0.31824763
##   5 0.000000000 0.00000000
##   6 0.077494996 0.26738480
##   7 0.000000000 0.00000000
## 
##    Soil_Type7
## Y           [,1]       [,2]
##   1 0.0000000000 0.00000000
##   2 0.0003839093 0.01958989
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0000000000 0.00000000
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
## 
##    Soil_Type8
## Y           [,1]       [,2]
##   1 0.0002007214 0.01416624
##   2 0.0004677516 0.02162256
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0000000000 0.00000000
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
## 
##    Soil_Type9
## Y           [,1]       [,2]
##   1 0.0007202357 0.02682762
##   2 0.0035478519 0.05945822
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0000000000 0.00000000
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
##
##    Soil_Type10
## Y          [,1]       [,2]
##   1 0.004374546 0.06599572
##   2 0.038112931 0.19146931
##   3 0.322246341 0.46734493
##   4 0.085145766 0.27916308
##   5 0.028878971 0.16747734
##   6 0.511366886 0.49988865
##   7 0.000000000 0.00000000
##
##    Soil_Type11
## Y          [,1]       [,2]
##   1 0.003494914 0.05901458
##   2 0.032111590 0.17629683
##   3 0.038057559 0.19133861
##   4 0.011105969 0.10482229
##   5 0.073116303 0.26034440
##   6 0.028452960 0.16626895
##   7 0.000000000 0.00000000
##
##    Soil_Type12
## Y         [,1]      [,2]
##   1 0.01266316 0.1118163
##   2 0.09649363 0.2952677
##   3 0.00000000 0.0000000
##   4 0.00000000 0.0000000
##   5 0.00000000 0.0000000
##   6 0.00000000 0.0000000
##   7 0.00000000 0.0000000
##
##    Soil_Type13
## Y           [,1]       [,2]
##   1 0.0106441386 0.10262019
##   2 0.0469737353 0.21158308
##   3 0.0013304390 0.03645155
##   4 0.0000000000 0.00000000
##   5 0.1367813074 0.34363888
##   6 0.0356734344 0.18548127
##   7 0.0003034901 0.01741885
##
##    Soil_Type14
## Y          [,1]       [,2]
##   1 0.000000000 0.00000000
##   2 0.000000000 0.00000000
##   3 0.003431132 0.05847631
##   4 0.054604350 0.22725891
##   5 0.000000000 0.00000000
##   6 0.018730340 0.13557591
##   7 0.000000000 0.00000000
##
##    Soil_Type15
## Y           [,1]       [,2]
##   1 0.0000000000 0.00000000
##   2 0.0000000000 0.00000000
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0000000000 0.00000000
##   6 0.0002144695 0.01464373
##   7 0.0000000000 0.00000000
##
##    Soil_Type16
## Y          [,1]       [,2]
##   1 0.002939978 0.05414196
##   2 0.006195503 0.07846748
##   3 0.003991317 0.06305177
##   4 0.020360944 0.14126433
##   5 0.003938041 0.06263424
##   6 0.014369460 0.11901257
##   7 0.000000000 0.00000000
##
##    Soil_Type17
## Y           [,1]       [,2]
##   1 0.0009799928 0.03128959
##   2 0.0033934056 0.05815415
##   3 0.0138645753 0.11693087
##   4 0.1675150393 0.37352150
##   5 0.0636650039 0.24417125
##   6 0.0401772948 0.19638186
##   7 0.0000000000 0.00000000
##
##    Soil_Type18
## Y           [,1]       [,2]
##   1 0.0002833714 0.01683130
##   2 0.0058071804 0.07598344
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0168023103 0.12853856
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
##
##    Soil_Type19
## Y           [,1]       [,2]
##   1 0.0115887100 0.10702560
##   2 0.0052953013 0.07257606
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0076135469 0.08692855
##   6 0.0000000000 0.00000000
##   7 0.0001820941 0.01349340
```

```
##
##    Soil_Type20
## Y           [,1]        [,2]
##   1 1.750409e-02 0.131140370
##   2 1.822466e-02 0.133763233
##   3 7.002311e-05 0.008367835
##   4 0.000000e+00 0.000000000
##   5 5.513258e-03 0.074051211
##   6 1.601373e-02 0.125532518
##   7 0.000000e+00 0.000000000
##
##    Soil_Type21
## Y           [,1]       [,2]
##   1 3.849128e-03 0.06192201
##   2 7.942952e-05 0.00891199
##   3 0.000000e+00 0.00000000
##   4 0.000000e+00 0.00000000
##   5 0.000000e+00 0.00000000
##   6 0.000000e+00 0.00000000
##   7 5.462822e-04 0.02336701
##
##    Soil_Type22
## Y          [,1]       [,2]
##   1 0.121944164 0.32722227
##   2 0.026485332 0.16057389
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.000000000 0.00000000
##   6 0.000000000 0.00000000
##   7 0.007405159 0.08573663
##
##    Soil_Type23
## Y          [,1]       [,2]
##   1 0.167395758 0.37332994
##   2 0.072995728 0.26013006
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.074691520 0.26291019
##   6 0.001715756 0.04138762
##   7 0.034294385 0.18198981
##
##    Soil_Type24
## Y          [,1]       [,2]
##   1 0.052594915 0.22322407
##   2 0.034225297 0.18180779
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.007088475 0.08389966
##   6 0.007935373 0.08872973
##   7 0.010257967 0.10076387
##
##    Soil_Type25
## Y           [,1]       [,2]
##   1 0.0005667428 0.02379968
##   2 0.0012488086 0.03531649
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0000000000 0.00000000
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
##
##    Soil_Type26
## Y           [,1]       [,2]
##   1 0.001286978 0.03585150
##   2 0.007678187 0.08728841
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.013783145 0.11659740
##   6 0.000000000 0.00000000
##   7 0.000000000 0.00000000
##
##    Soil_Type27
## Y           [,1]       [,2]
##   1 0.002951786 0.05425025
##   2 0.001584178 0.03977028
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.000000000 0.00000000
##   6 0.000000000 0.00000000
##   7 0.001456753 0.03814078
##
##    Soil_Type28
## Y            [,1]       [,2]
##   1 0.0002007214 0.01416624
##   2 0.0031507043 0.05604276
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0013126805 0.03620952
##   6 0.0000000000 0.00000000
##   7 0.0000000000 0.00000000
##
##    Soil_Type29
## Y          [,1]      [,2]
##   1 0.19836589 0.3987704
##   2 0.25085607 0.4335068
##   3 0.00000000 0.0000000
##   4 0.00000000 0.0000000
##   5 0.11682856 0.3212370
##   6 0.00000000 0.0000000
##   7 0.03878604 0.1930905
##
##    Soil_Type30
## Y          [,1]      [,2]
##   1 0.03576383 0.1857013
##   2 0.07142038 0.2575263
##   3 0.00000000 0.0000000
##   4 0.00000000 0.0000000
##   5 0.22223681 0.4157767
##   6 0.00000000 0.0000000
##   7 0.01019727 0.1004684
##
##    Soil_Type31
## Y          [,1]       [,2]
##   1 0.056544404 0.23097067
##   2 0.046748685 0.21110055
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.033342085 0.17954003
##   6 0.003860452 0.06201471
##   7 0.011047041 0.10452592
##
##    Soil_Type32
## Y          [,1]      [,2]
##   1 0.100425647 0.3005676
##   2 0.105005825 0.3065616
##   3 0.003081017 0.0554223
##   4 0.000000000 0.0000000
##   5 0.048962982 0.2158048
##   6 0.011438376 0.1063407
##   7 0.041517451 0.1994898
##
##    Soil_Type33
## Y           [,1]       [,2]
##   1 0.0853656377 0.27942585
##   2 0.0892831927 0.28515270
##   3 0.0001750578 0.01323001
##   4 0.0000000000 0.00000000
##   5 0.0526384878 0.22332538
##   6 0.0320274521 0.17607927
##   7 0.0306525038 0.17237961
##
##    Soil_Type34
## Y           [,1]       [,2]
##   1 0.0004191535 0.02046901
##   2 0.0051143785 0.07133193
##   3 0.0000000000 0.00000000
##   4 0.0000000000 0.00000000
##   5 0.0018377527 0.04283242
##   6 0.0008578782 0.02927804
##   7 0.0026100152 0.05102314
##
##    Soil_Type35
## Y           [,1]       [,2]
##   1 4.268282e-03 0.06519270
##   2 4.854026e-05 0.00696693
##   3 0.000000e+00 0.00000000
##   4 0.000000e+00 0.00000000
##   5 0.000000e+00 0.00000000
##   6 0.000000e+00 0.00000000
##   7 4.552352e-02 0.20845567
##
##    Soil_Type36
## Y            [,1]        [,2]
##   1 6.493928e-05 0.008058253
##   2 1.588590e-04 0.012602956
##   3 0.000000e+00 0.000000000
##   4 0.000000e+00 0.000000000
##   5 0.000000e+00 0.000000000
##   6 0.000000e+00 0.000000000
##   7 3.277693e-03 0.057158976
##
##    Soil_Type37
## Y          [,1]      [,2]
##   1 0.00000000 0.0000000
##   2 0.00000000 0.0000000
##   3 0.00000000 0.0000000
##   4 0.00000000 0.0000000
##   5 0.00000000 0.0000000
##   6 0.00000000 0.0000000
##   7 0.01329287 0.1145293
##
##    Soil_Type38
## Y          [,1]       [,2]
##   1 0.04149620 0.19943546
##   2 0.00259911 0.05091529
##   3 0.00000000 0.00000000
##   4 0.00000000 0.00000000
##   5 0.00000000 0.00000000
##   6 0.00000000 0.00000000
##   7 0.29669196 0.45681343
##
##    Soil_Type39
## Y           [,1]       [,2]
##   1 0.037399123 0.18973835
##   2 0.001248809 0.03531649
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.000000000 0.00000000
##   6 0.000000000 0.00000000
##   7 0.272534143 0.44527668
##
##    Soil_Type40
## Y           [,1]       [,2]
##   1 0.022852724 0.14943429
##   2 0.001156141 0.03398248
##   3 0.000000000 0.00000000
##   4 0.000000000 0.00000000
##   5 0.000000000 0.00000000
##   6 0.000000000 0.00000000
##   7 0.175477997 0.38038698
```

The output above shows that it learned the following correlations and probabilities:

The prior for the seven Cover Types is shown as roughly: 36% for cover type 1, 49% for cover type 2, 6% for cover type 3, .5% for cover type 4, 1.6% for cover type 5, and 3% for cover type 6. This shows that cover type 1 and 2 are the most likely overall by a significant amount.

For the predictors: You can tell they are all continuous probabilities (don't sum to 1, instead they will give us the mean for each).

Elevation: Here we can see that the means for elevation are all very similar and therefore don't tell us much by themselves. Some are a slightly higher such as for cover types 1, 2, 5, and 7, while the others are slightly lower. But overall there isn't a significant difference.

Aspect: Here we can see that the means for aspect are all very similar and therefore don't tell us much by themselves. Some are a slightly higher such as for cover types 3 and 6, with some slightly lower such as for cover types 4 and 5, but the rest are nearly the same. Overall there isn't a significant difference.

Slope: Here we can see that the means for slope are also all very similar and therefore don't tell us much by themselves. Some are a slightly higher such as for cover types 3, 4, and 6, with some slightly lower such as for cover types 1 and 2, and cover types 5 and 7 in the middle. Overall there isn't a significant difference.

We see this pattern continue in the rest of the predictors except the following:

Horizontal Distance to Roadways: In this one we can see that the means for cover types 1, 2, and 7 are about double compared to the rest of the means. This could indicate that the father away from a road way, the more likely it is to be cover type 1, 2, or 7. Also implies the closer to a road way, the more likely it is to be of the other cover types (3, 4 5, and 6).

Horizontal Distance to Fire Points: Here we see again that the means for cover types 1, 2, and 7 are about double compared to the rest of the means. This could indicate that the father away from a fire point, the more likely it is to be cover type 1, 2, or 7. Also implies the closer to a fire point, the more likely it is to be of the other cover types (3, 4 5, and 6). It can also imply that fire points and road ways are highly correlated with one another.

Wilderness Area 1: The means for cover types 3, 4, and 6 is 0, showing it is highly unlikely for those cover types to be in wilderness area 1. Cover types 1, 2, and 3 have means close to about .5 whereas cover type 7 has a mean of about .25 this shows that cover types 1, 2, and 3, are very likely, cover type 7 is about half as likely, and cover types 3, 4, 5, and 6 are very unlikely to be correlated with wilderness area 1.

Wilderness Area 2: The means for cover types 3, 4, 5, and 6 is 0, showing it is highly unlikely for those cover types to be in wilderness area 2. Cover type 7 has a mean of about .1, cover type 1 has a mean of about .09, and cover type 2 has a mean of about .032. This shows that cover types 1 and 7 are about 3 times as likely as cover type 2, and cover types 3, 4, 5, and 6 are highly unlikely to be correlated with wilderness area 2.

Wilderness Area 3: The mean for cover type 4 is 0, showing it is highly unlikely for cover type 4 to be in wilderness area 3. Cover types 5 and 7 have means of about .63, and the rest of the cover types 1, 2, 3, and 6 have means of about .42. This shows that cover types 5 and 7 are more likely than cover types 1, 2, 3, and 6, and cover type 4 is highly unlikely to be correlated with wilderness area 3.

Wilderness Area 4: The means for cover types 1, 4, 5, and 7 is 0, showing it is highly unlikely for those cover types to be in wilderness area 4. Cover types 3 and 6 have means of about .58, and cover type 2 has a mean of about .01. This shows that cover types 31 and 6 are much more likely than cover type 2, and cover types 1, 4, 5, and 7 are highly unlikely to be correlated with wilderness area 4.

The soil types have very similar patterns, however there are 40 columns and it would be very lengthy to review all of them in this summary.

Evaluate on the test data

```
pbayes <- predict(nb1, newdata=test, type="class")
print(paste("Mean: ", mean(pbayes==test$Cover_Type)))
```

```
## [1] "Mean:  0.0701875166734077"
```

```
pbayes_raw <- predict(nb1, newdata=test, type="raw")
head(pbayes_raw)
```

```
##                 1            2            3            4            5
## [1,] 9.332138e-76 3.206255e-43 6.923390e-45 9.999889e-01 1.056065e-40
## [2,] 3.135398e-10 1.554618e-49 3.171607e-01 6.568027e-01 2.603633e-02
## [3,] 2.007640e-10 1.026520e-49 3.384823e-01 6.535932e-01 7.924173e-03
## [4,] 8.682481e-11 4.067058e-50 3.314763e-01 6.635681e-01 4.955285e-03
## [5,] 1.826927e-10 8.896876e-50 3.431334e-01 6.376655e-01 1.920098e-02
## [6,] 2.541663e-13 2.667511e-22 9.166345e-19 7.907062e-40 9.999755e-01
##                 6            7
## [1,] 4.007004e-12 1.112515e-05
## [2,] 2.251117e-07 1.712116e-08
## [3,] 3.233375e-07 4.733661e-09
## [4,] 3.124940e-07 2.707100e-09
## [5,] 1.588749e-07 1.357396e-08
## [6,] 3.774442e-24 2.447682e-05
```

```
confusionMatrix(data=pbayes, reference=test$Cover_Type)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2     3     4     5     6     7
##          1    45   210     0     0     0     0     0
##          2     0     0     0     0     0     0     0
##          3 15518 19381  2776    10   632  1201   162
##          4   280  4627  4404   576   253  2118     0
##          5 11529 23483    12     0   990    58   106
##          6     8    51     0     0     0     2     0
##          7 15071  8933     0     0     0     0  3767
##
## Overall Statistics
##
##                Accuracy : 0.0702
##                  95% CI : (0.0687, 0.0717)
##     No Information Rate : 0.4878
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0357
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                       Class: 1 Class: 2 Class: 3 Class: 4 Class: 5  Class: 6
## Sensitivity          0.0010600   0.0000  0.38598 0.982935  0.52800 5.919e-04
## Specificity          0.9971526   1.0000  0.66147 0.898959  0.69222 9.995e-01
## Pos Pred Value       0.1764706      NaN  0.06996 0.046990  0.02736 3.279e-02
## Neg Pred Value       0.6342671   0.5122  0.94229 0.999904  0.98894 9.709e-01
## Prevalence           0.3653176   0.4878  0.06189 0.005043  0.01614 2.908e-02
## Detection Rate       0.0003873   0.0000  0.02389 0.004957  0.00852 1.721e-05
## Detection Prevalence 0.0021944   0.0000  0.34147 0.105488  0.31133 5.249e-04
## Balanced Accuracy    0.4991063   0.5000  0.52372 0.940947  0.61011 5.000e-01
##                      Class: 7
## Sensitivity           0.93358
## Specificity           0.78600
## Pos Pred Value        0.13565
## Neg Pred Value        0.99697
## Prevalence            0.03472
## Detection Rate        0.03242
## Detection Prevalence  0.23899
## Balanced Accuracy     0.85979
```

As seen above, our accuracy for bayes model is about 7%, which is significantly less than our logical regression model.We think a reason why is because of large our data set is, the bayes algorithm tends to perform better on smaller datasets than large ones such as this one. We also think that the means and variance for this dataset isn't super helpful for predictions when compared to a linear algorithm. The bayes class independence assumption causes it to perform very poorly here, as all the classes in our dataset are very much connected/dependent on one another, such as soil type, wilderness area, and elevation.

Classification Metrics:

Accuracy:

As shown from above, the overall accuracy our Logistic Regression Models was 90.9%. Whereas our accuracy from our Native Bayes Model was 7%. According to this metric our logistic regression models are way more accurate than our bayes model.

Sensitivity and Specificity:

Sensitivity: True Positive Rate, (TP / (TP+FN)) As you can see from above our overall sensitivity rate is .96 for our One vs. All Logistic Regressions. For the bayes model it was .04. However, for classes 4 and 7 the rate was very high, meaning it was giving true positives very well. What this tells is is the true positive rate, so our logistic regression models had a very high true positive rate when predicting. The Bayes model did better with specific cover types but was overall worse at predicting true positives than our regression models.

Specificity: True Negative Rate, (TN / (TN+FP)) «««< HEAD As you can see from above our overall specificity rate is .34 for our One vs. All Logistic Regressions. For the bayes model it was _____. What this tells is is the true negative rate, so our logistic regression models had a very low true negative rate when predicting. ======= As you can see from above our overall specificity rate is .34 for our One vs. All Logistic Regressions. For the bayes model it was (overall of all classes) is .4 which is slightly higher than the logistic regressions. The Bayes model did better with specific cover types and was overall slightly better at predicting true negatives than our regression models. »»»> b28417f772ccb35f26efecd191f5e7369306be33

Kappa:

Kappa attempts to adjust accuracy by accounting for correct prediction by chance. Our overall Kappa for our one vs. all logistic regressions is .36 which is a fair agreement. This means is better than chance, but not quite a good agreement either. For bayes it was .86 which is very good agreement, meaning that the model's predictions are better than chance. (Guessing wouldn't be equivalent) For classes 1, 2, 4, and 6 it was practically 1 meaning that the true negative for bayes is very good. The bayes model performed much better in for this metric than our logistic regression one vs all models.

ROC Curves and AUC:

This metric measures how well our model fits the data according to the true positive and false positive rates. Ideally the graph should be almost as square as possible to the y-axis until it tops off in a straight line. For our logistic regression one for all models by looking at these graphs and the area under the curve (AUC) we can conclude: - Cover Type 1: The ROC graph is not very square and has a fair amount of curving in the top left. The AUC is .84 which is still considered excellent. - Cover Type 2: The ROC graph is not very square and has a fair amount of curving in the top left (slightly more than cover type 1). The AUC is .83 which is still considered excellent. - Cover Type 3: The ROC graph is very square and has almost no curving in the top left. The AUC is .98 which is considered outstanding. - Cover Type 4: The ROC graph is completely square and has no curving in the top left. The AUC is 1 which is considered outstanding. - Cover Type 5: The ROC graph is not square and has a lot of curving in the beginning and even starts off down and then moves up into a smooth curve. The AUC is .71 which is still considered acceptable. - Cover Type 6: The ROC graph is fairly square and has a little bit of curving in the top left. The AUC is .96 which is considered outstanding. - Cover Type 7: The ROC graph is very square and has almost no curving in the top left. The AUC is .98 which is considered outstanding.

Naive Bayes vs Logistic Regression

Logistic regression is a very powerful tool. Not only is it a good way to classify data that is relatively low cost, but it is fairly accurate when classes are linearly separable and provides a nice probabilistic output. It can underfit data that is too complex for it to be able to split cleanly, meaning that it might not be as accurate with data sets that have high variance. Naive Bayes works well with small data sets, is easy to implement, easy to understand, and can handle high dimensions as well. However, guesses are made for values in the test set that aren't present in the training set, can't perform as well when predictors are not independent, and can be outperformed when used on larger data sets.

Classification Metrics: Benefits/Drawbacks

Kappa tries to adjust the accuracy rate given to a model by taking into account the chance that some of the model's accuracy comes from chance alone. It assess how the actual and expected agreement compare to one another. The higher the number calculated, the better it is. It's very good to test the reliability of a model. ROC is a visualization of the performance of an algorithm and demonstrates the tradeoffs for predicting more true positives while avoiding false negatives. AUC shows us the area underneath the curve of an ROC diagram.MCC measures the differences between the predicted values and actual values for a model. Overall these metrics are good for getting a better understanding of how a algorithm works and the ways it could be improved, but they tend to be slow and can be difficult to implement.