Team Checkpoint 1

Forest Cover Team B

Annie Condon, Matt Hayden, Matt Robertson, Yvette Gonzalez

Contents

1 Introduction

The U.S. Forest Service relies on an accurate understanding of its forests composition in order to best protect and manage the forest land. Conducting accurate inventory of forest composition by direct observation or remotely sensed data is often too expensive and time consuming to do at large-scale. Predictive analytics can be employed to use the results of a small-scale survey to create a model that can be applied across a large region, using descriptive features extracted from maps of the area.

In this paper, our objective is to predict the forest cover type given a set of cartographic features and a variety of mulitclass classification models. Our models will be evaluated using predictive accuracy.

2 The Modeling Problem

Our modeling problem is to predict the forest cover type as a multiclass classification problem based on the associated features. A multiclass classification problem classifies instances into one of the more than two classes. Our forest cover type is defined as one of seven, mutually exclusive, forest cover type classes, shown in table 1 below.

Table 1: Cover Type Classes

Cover Type	Number of Observations
Spruce/Fir	211840
Lodgepole Pine	283301
Ponderosa Pine	35754
Cottonwood/Willow	2747
Aspen	9493
Douglas/Fir	17367
Krummholz	20510

Several algorithms have been developed to solve multiclass classification problems. The possible algorithms we will consider for our problem are: neural networks, k-nearest neighbors, random forest and support vector machines.

We will evaluate our models using classification accuracy.

3 The Data

Our data set consists of 581,012 observations of the 30 x 30 meter cells of forest and 54 features associated with each cell. The features are derived from 12 attributes, with area and soil type binarized so that there are 4 binary area designators and 40 binary soil type designators. There is no missing data. The feature descriptions are listed in the table below:

Table 2: Features

Feature	Descriptions
Elevation	Elevation in meters
Aspect	Aspect in degrees aziumuth
Slope	Slope in degrees
HDist.Hydrology	Horizontal distance to nearest surface water feature in meters
VDist.Hydrology	Vertical distance to nearest surface water feature in meters

Feature	Descriptions
HDist.Roadway	Horizontal distance to nearest roadway in meters
Hillshade.9am	Hillshade index at 9am, summer solstice
Hillshade.12pm	Hillshade index at noon, summer soltice
Hillshade.3pm	Hillshade index at 3pm, summer solstice
HHDist.FirePoint	Horizontal Distance to nearest wildfire ignition points
Area	Wilderness area designation - 4 binary areas
SoilType	Soil Type designation - 40 binary values

3.1 Response Variable

Figure one shows the frequency of each class of cover type in our data set. 85 percent of the observations fall into classes 1 and 2, Spruce/Fir and Lodgepole Pine. Class 4, Cottonwood/Willow has the least amount of observations at 2,747.

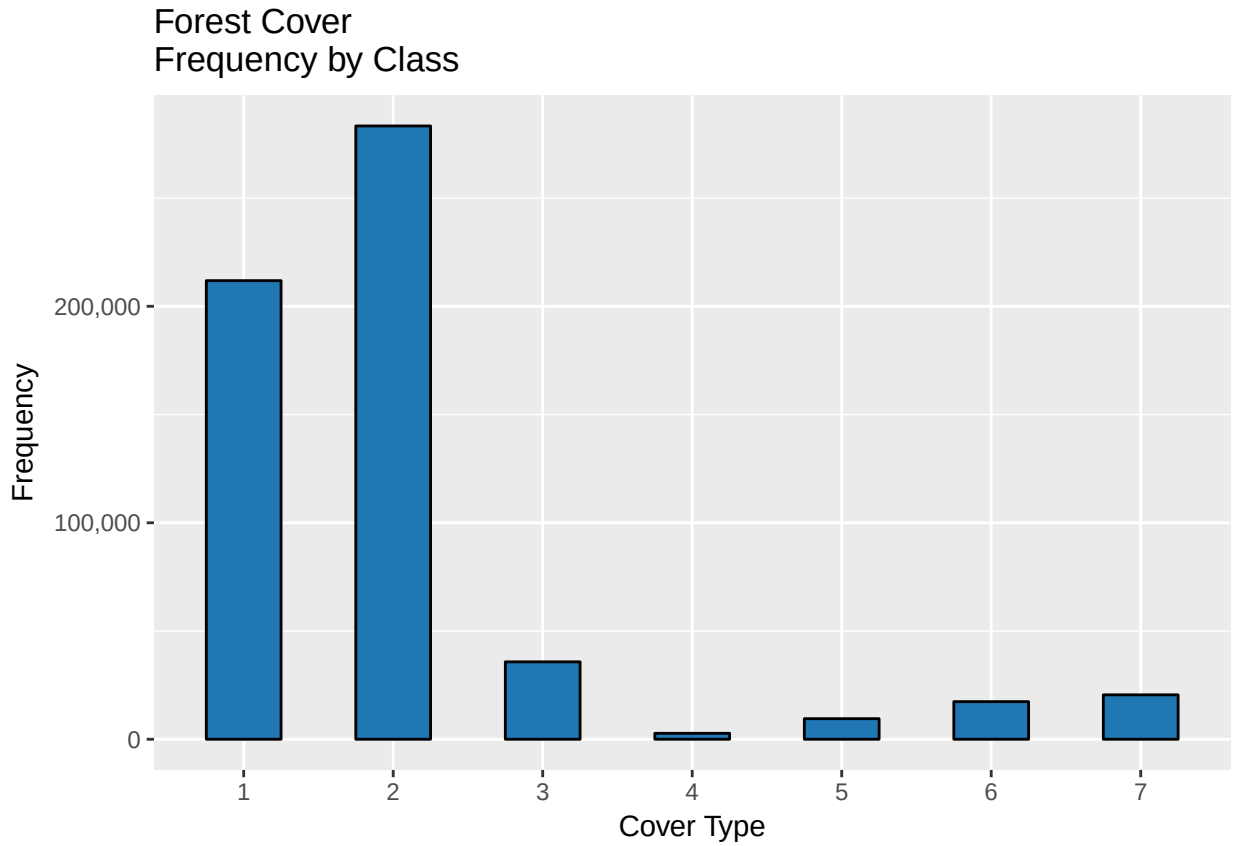


Figure 1: Frequency of each Cover Type Class

3.2 Continuous Variables

Table 3 includes some standard statistical measures of central tendency and variation for our continuous variables, for investigation of unusual values. We would first like to note that an Aspect value of 0 and 360 would both be equal to true north, so we should standardize that value. Next, we will point out that while a negative distance value for VDist.Hydrology may appear to be an error, it is in fact reasonable due to it

being a vertical measurement from differing altitudes, making a negative distance possible. Our Hillshade values fall into the hillshade index integer value range of 0 to 255. Also, our horizontal distances all have minimum values of 0.

Table 3: Summary Statistics for Continuous Data

	mean	sd	median	min	max	nmiss	type
Elevation	2959.3653	279.9847	2996	1859	3858	0	Continuous
Aspect	155.6568	111.9137	127	0	360	0	Continuous
Slope	14.1037	7.4882	13	0	66	0	Continuous
HDist.Hydrology	269.4282	212.5494	218	0	1397	0	Continuous
VDist.Hydrology	46.4189	58.2952	30	-173	601	0	Continuous
HDist.Roadway	2350.1466	1559.2549	1997	0	7117	0	Continuous
Hillshade.9am	212.146	26.7699	218	0	254	0	Continuous
Hillshade.12pm	223.3187	19.7687	226	0	254	0	Continuous
Hillshade.3pm	142.5283	38.2745	143	0	254	0	Continuous
HDist.FirePoint	1980.2912	1324.1952	1710	0	7173	0	Continuous

3.3 Binary Variables

Our data's binary variables consist of 4 different wilderness area designators and 40 different soil type designators. Table 4 and 5 below represent the frequency of each of the areas and soil types, in descending order.

Table 4: Area Type Counts

Name	Count
Area1	260796
Area3	253364
Area4	36968
Area2	29884

Table 5: Soil Type Counts

Name	Count
SoilType29	115247
SoilType23	57752
SoilType32	52519
SoilType33	45154
SoilType22	33373
SoilType10	32634
SoilType30	30170
SoilType12	29971
SoilType31	25666
SoilType24	21278
SoilType13	17431
SoilType38	15573
SoilType39	13806
SoilType11	12410
SoilType4	12396
SoilType20	9259

Name	Count
SoilType40	8750
SoilType2	7525
SoilType6	6575
SoilType3	4823
SoilType19	4021
SoilType17	3422
SoilType1	3031
SoilType16	2845
SoilType26	2589
SoilType18	1899
SoilType35	1891
SoilType34	1611
SoilType5	1597
SoilType9	1147
SoilType27	1086
SoilType28	946
SoilType21	838
SoilType14	599
SoilType25	474
SoilType37	298
SoilType8	179
SoilType36	119
SoilType7	105
SoilType15	3

4 Exploratory Data Analysis

Our next step is to explore the relationships in our data. We will begin by looking at boxplots of our scaled continuous variables in order to understand their relative distribution. We note that Hillshade.12pm and VDist.Hydrology have more pronounced skews than the other variables. We will need to investigate transformations if we use modeling techniques that can be negatively effected by outliers.

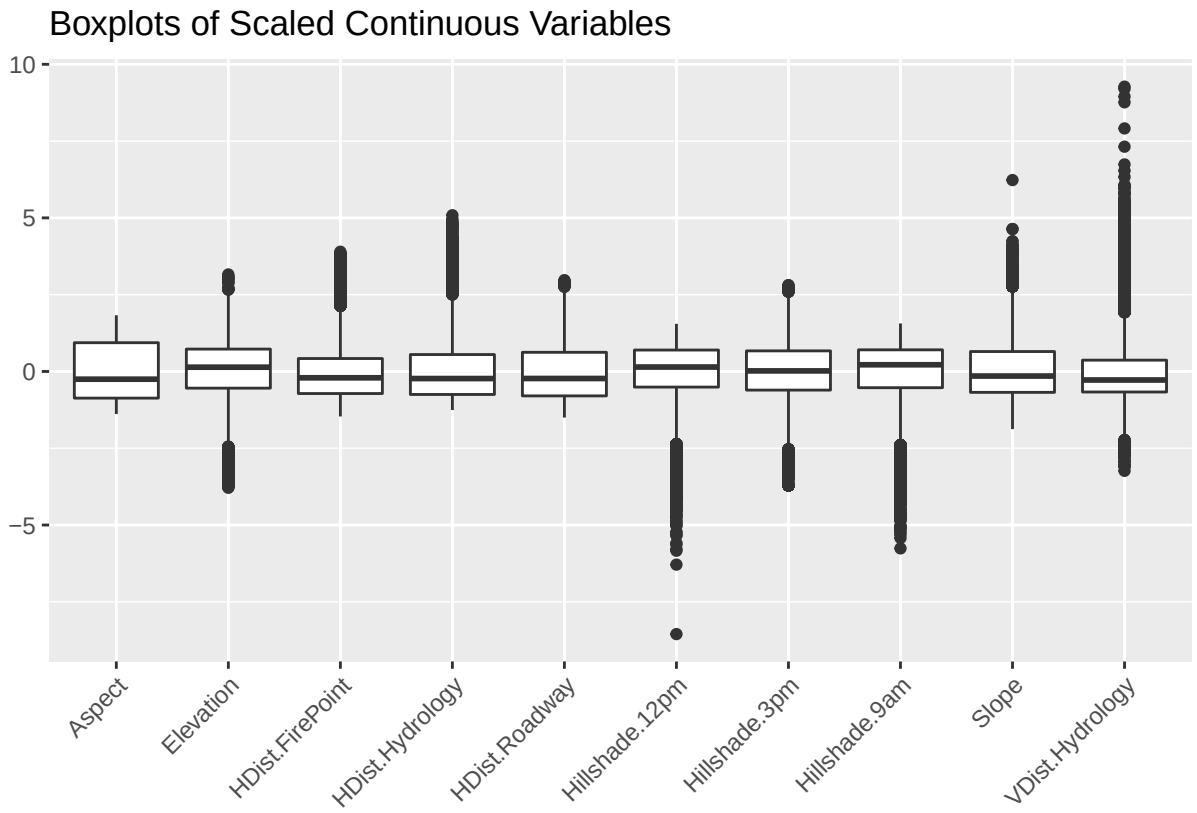


Figure 2: Boxplots of Scaled Continuous Variables

Our data exploration is informed by our statistical problem, which is one of multiclass classification. We will therefore be interested in exploring our predictors by each of our classes. The density plots below superimpose the density estimates for each variable by class, 1-7. We note that our variable Aspect shows multimodel distributions, indicating that it may have more than one grouping included. This is due to the value of 0 and 360 both being equal to true north, which we will have to address in data preparation. We will also note that several of our variables show distinct distributions by class, indicating that they would serve as a good predictor.

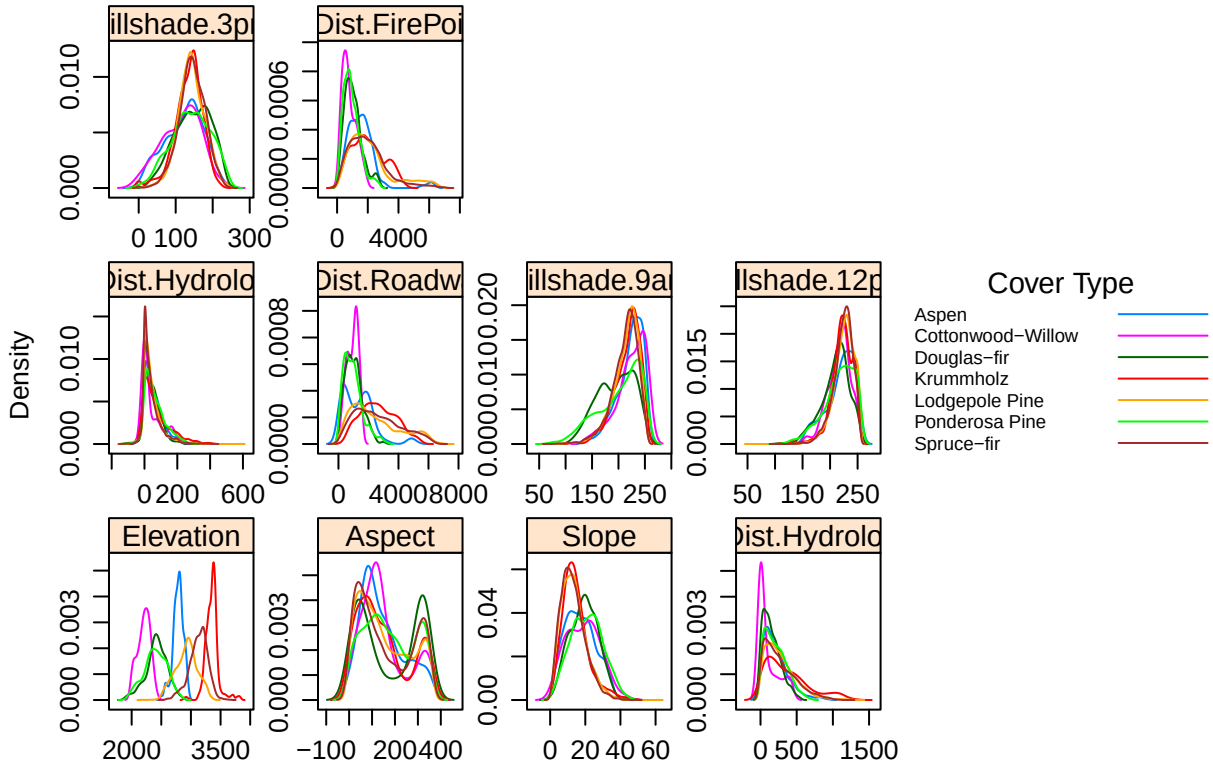


Figure 3: Density Plots of Continuous Variables by Class

Figure 4 shows density plots by each of the four wilderness areas.Several variable values appear to be distinguishable by wilderness area.

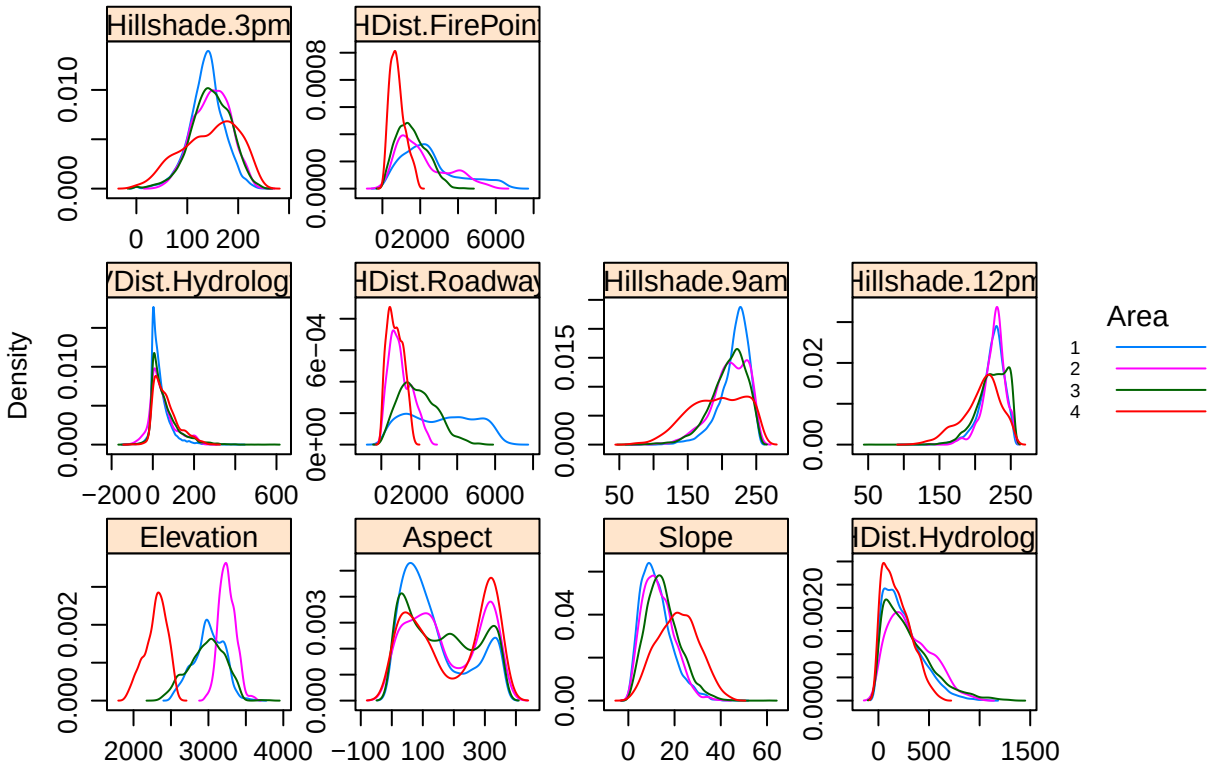


Figure 4: Density Plots of Continuous Variables by Area

Next, we will look at the correlations between the our predictor variables. Highly correlated predictors can have a negative effect on some of our modeling algorithms. Below is a correlation matrix with dark blue colors indicating a strong positive correlation and dark red colors indicating a strong negative correlation. We see the following variables with higher correlations that should be investigated further:

High Positive Collinearity

- VDist.Hydrology and HDist.Hydrology
- Aspect and Hillshade.3pm
- Hillshade.3pm and Hillshade.12pm

High Negative Collinearity

- Aspect and Hillshade.9am
- Slope and Hillshade.12pm
- Hillshade.3pm and Hillshade.9am

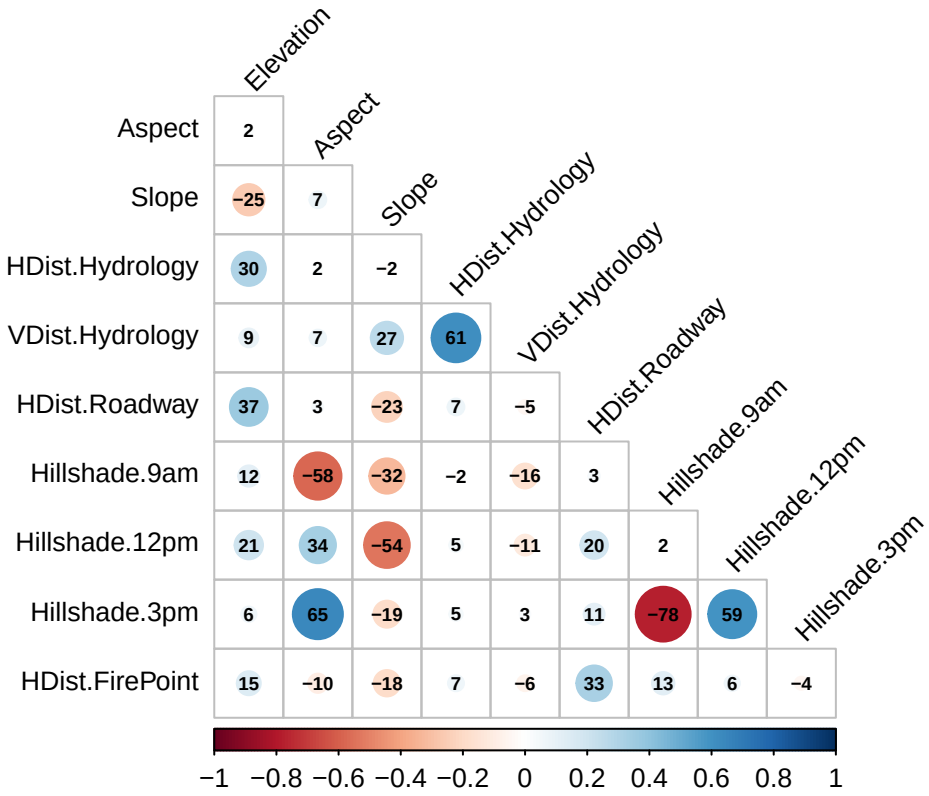


Figure 5: Correlation Matrix

Figure 6 shows a barchart of our 40 different soil types which shows the proportion of frequency of each class. We can see that soil types 1-7 have similar proportions, as do 19-33 and 38-40. A few soil types have only one cover type class, making them especially good predictors.

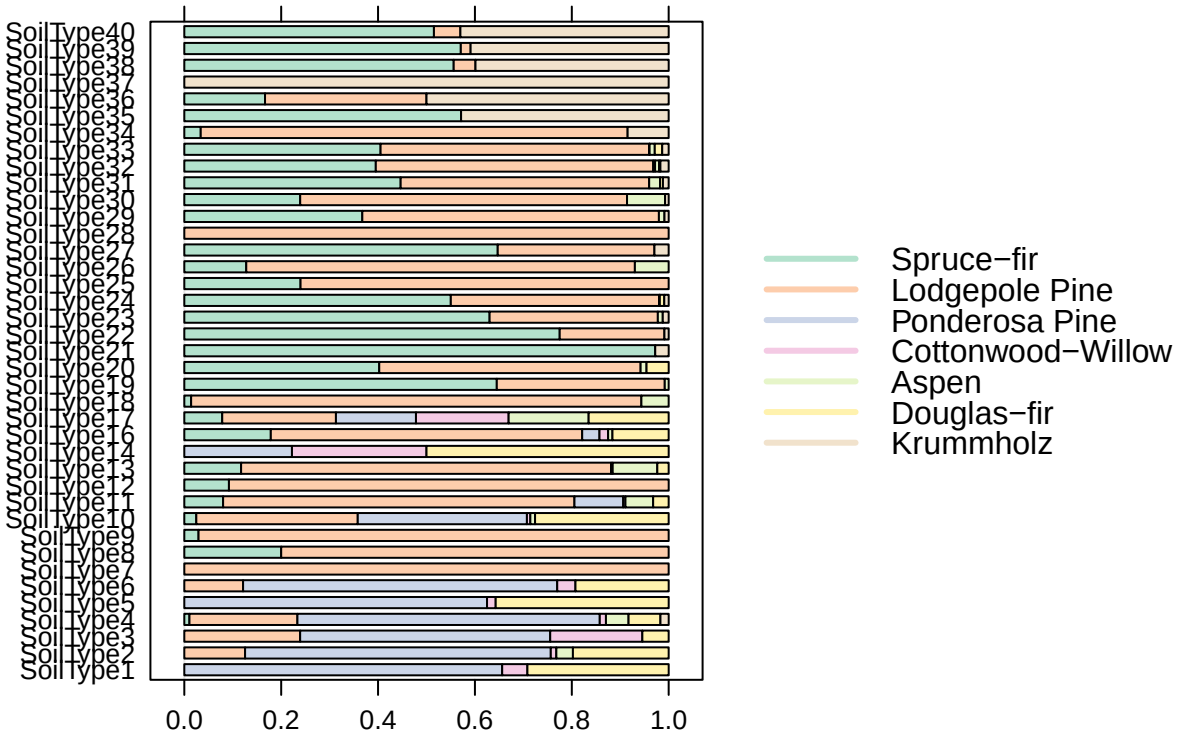


Figure 6: Soil Barchart

Figure 7 shows a barchart of our 4 different wilderness areas with the proportion of frequency of each class. We can see that the class composition in area 4 is distinguishable from the other areas, making it a good predictor.

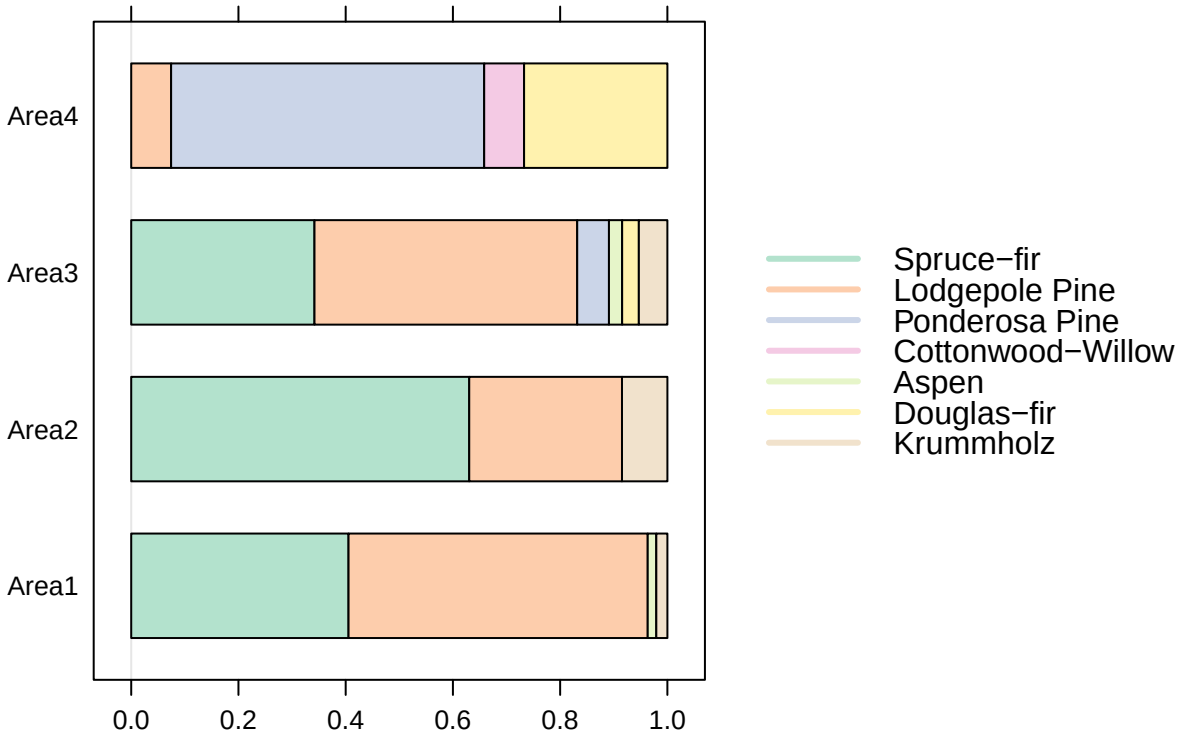


Figure 7: Area Barchart

Next, we look at the horizontal distance to the roadway, by class. We can see that most observations in classes 5-7, Aspen, Cottonwood-Willow and Douglas-fir are closest to the roadways.

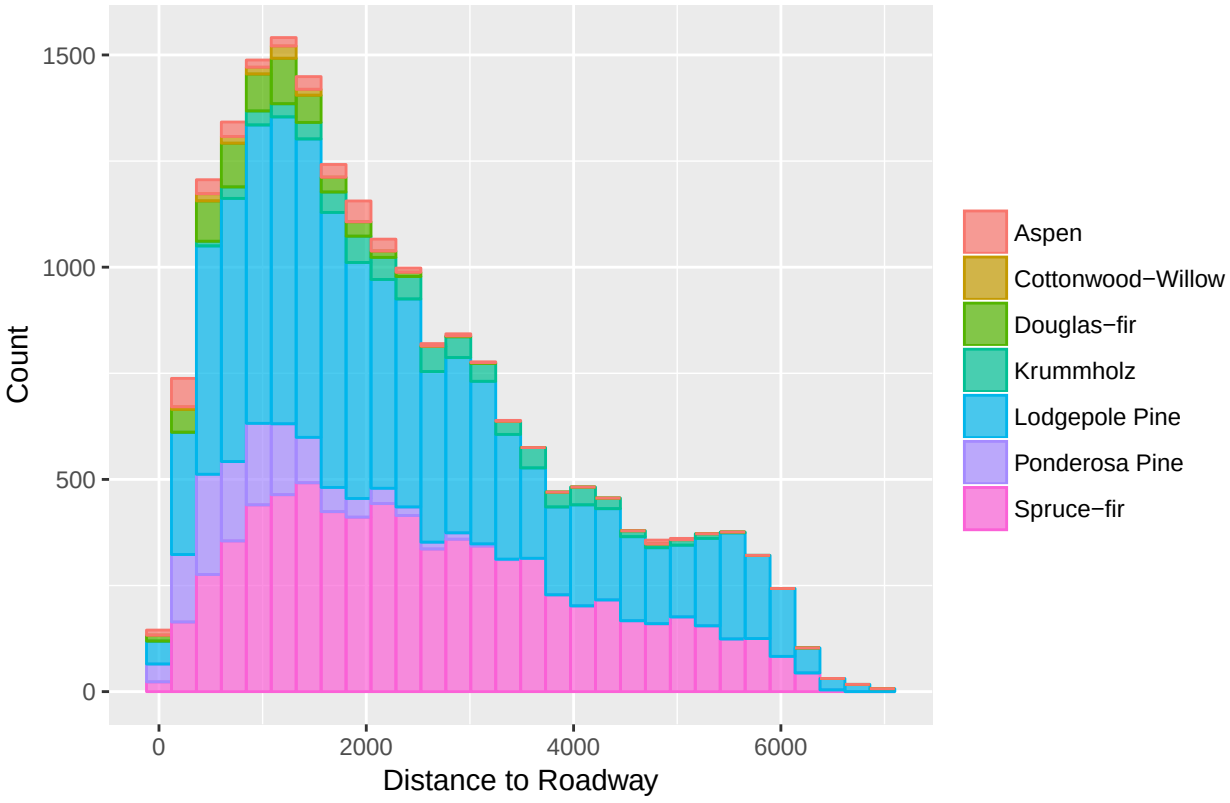
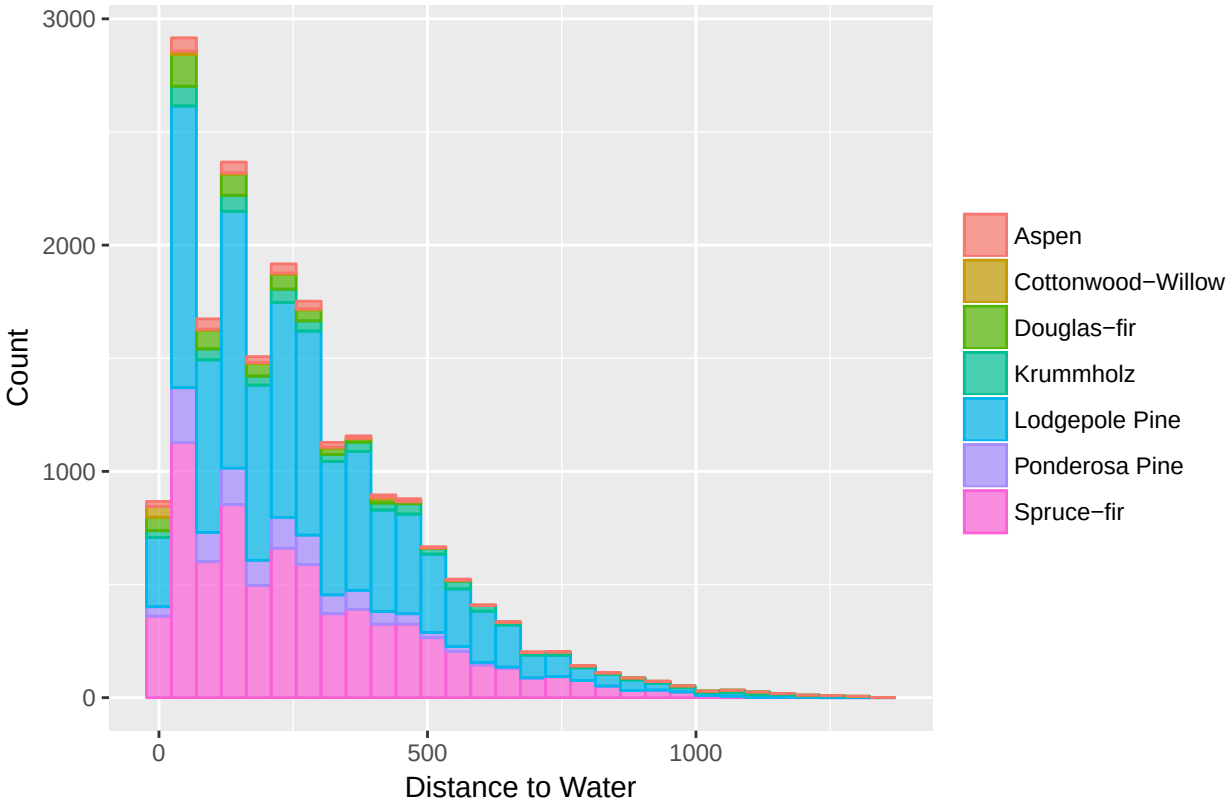


Figure 9 shows horizontal distance to the water, by class. Again, we see that most observations in classes 5-7 are closest to a water source.



5 Conclusion

Our initial analysis for our forest cover type prediction problem included defining our modeling problem, doing a data quality and inventory check and performing a preliminary exploratory data analysis. The results of our data quality check showed that we had no missing data and gave us a cursory understanding of our data set.

Our preliminary exploratory analysis revealed some potentially useful predictors and helped us to understand the relationships in our data. Our next step will be to build a few exploratory models in order to gain further insights into predictors we can use in our modeling and to prepare our data for modeling.

Team Checkpoint 1 R Code

Appendix

Annie Condon, Matt Hayden, Matt Robertson, Yvette Gonzalez

```
# read in the data, create dataframe
gz = gzfile('data/covtype.data.gz','rt')
forest.orig = read.csv(gz,header=F)
forest.orig.colnames = t(read.csv('data/covtyp.colnames.csv',header=F))
colnames(forest.orig) = forest.orig.colnames

# identify continuous variables
forest.var.continuous = c('Elevation','Aspect','Slope', 'HDist.Hydrology', 'VDist.Hydrology',
                          'HDist.Roadway', 'Hillshade.9am', 'Hillshade.12pm', 'Hillshade.3pm',
                          'HDist.FirePoint')

# for speed, will perform eda on subset until ready to do a full run
set.seed(33)
forest = forest.orig[sample(nrow(forest.orig),20000),]
forest.var.discrete.indices = grep("^Area|^SoilType|CoverType", colnames(forest))
forest[,forest.var.discrete.indices] = as.factor(unlist(forest[,forest.var.discrete.indices]))

forest.numeric = as.data.frame(sapply(forest,as.numeric))

covertype.names = c('Spruce-fir','Lodgepole Pine','Ponderosa Pine','Cottonwood-Willow','Aspen','Douglas

forest$CoverType[forest$CoverType==1] = 'Spruce-fir'
forest$CoverType[forest$CoverType==2] = 'Lodgepole Pine'
forest$CoverType[forest$CoverType==3] = 'Ponderosa Pine'
forest$CoverType[forest$CoverType==4] = 'Cottonwood-Willow'
forest$CoverType[forest$CoverType==5] = 'Aspen'
forest$CoverType[forest$CoverType==6] = 'Douglas-fir'
forest$CoverType[forest$CoverType==7] = 'Krummholz'


# add Area column
# are any in multiple areas? NO. all belong to only one area
idx = grep("Area", colnames(forest))
temp = forest.numeric[,idx]
temp.rows = temp[apply(temp,1,sum) > 1,]
temp$Z.Area = apply(temp,1,function(x) {
  return(which.max(x))
})
forest.numeric$Z.Area = temp$Z.Area

forest.scaled = as.data.frame(scale(forest.numeric))

library(lattice)
library(ggplot2)
library(corrplot)
library(MASS)
```

```
cover_types <- c("Spruce/Fir", "Lodgepole Pine", "Ponderosa Pine", "Cottonwood/Willow", "Aspen",
                 "Douglas/Fir", "Krummholz")
instances <- c(211840, 283301, 35754, 2747, 9493, 17367, 20510)
ct.df <- data.frame(cover_types, instances)
colnames(ct.df) <- c("Cover Type", "Number of Observations")
```

```
knitr::kable(ct.df, caption = "Cover Type Classes", format="pandoc")
```

```
features <- c("Elevation", "Aspect", "Slope", "HDist.Hydrology", "VDist.Hydrology",
              "HDist.Roadway", "Hillshade.9am", "Hillshade.12pm", "Hillshade.3pm",
              "HHDist.FirePoint", "Area", "SoilType")
descriptions <- c("Elevation in meters", "Aspect in degrees aziumuth", "Slope in degrees",
                  "Horizontal distance to nearest surface water feature in meters",
                  "Vertical distance to nearest surface water feature in meters",
                  "Horizontal distance to nearest roadway in meters", "Hillshade index at 9am, summer sol
                 "Hillshade index at noon, summer soltice", "Hillshade index at 3pm, summer solstice",
              "Horizontal Distance to nearest wildfire ignition points", "Wilderness area designation -
              "Soil Type designation - 40 binary values")
features.df <- data.frame(features, descriptions)
colnames(features.df) <- c("Feature", "Descriptions")
```

```
knitr::kable(features.df, caption = "Features", format="pandoc")
```

```
library(scales)

ggplot(as.data.frame(table(forest.orig$CoverType)), aes(x=Var1, y = Freq)) + ggtitle("Forest Cover
Frequency by Class") + geom_bar(stat = "identity", fill="#1f78b4", width=.5,
                                color="black") + xlab("Cover Type") + scale_y_continuous(name="Frequency
```

```
options(digits=3)
my.summary <- function(x,...){
  c(mean=round(mean(x, ...), digits = 4),
    sd=round(sd(x, ...), digits=4),
    median=median(x, ...),
    min=min(x, ...),
    max=max(x,...),
    nmiss=sum(is.na(x,...)),
    type="Continuous")
}


forest.stats= apply(forest.orig[,1:10], 2, my.summary)

library(knitr)
kable(t(forest.stats), caption= "Summary Statistics for Continuous Data", format="pandoc")
```

```
## area
forest.area= forest.orig[11:14]

summary.area <- data.frame(
  Name = character(),
  Count = numeric(),
  stringsAsFactors = F)

for (i in 1:4){
  summary.area[i,1] <- names(forest.area[i])
  summary.area[i,2] <- sum(forest.area[,i])
}

area<-summary.area[with(summary.area,order(-Count)),]
kable(area,caption= "Area Type Counts", format="pandoc", row.names = F)
```

```
## soil
forest.soil= forest.orig[15:54]

summary.soil <- data.frame(
  Name = character(),
  Count = numeric(),
  stringsAsFactors = F)

for (i in 1:40){
  summary.soil[i,1] <- names(forest.soil[i])
  summary.soil[i,2] <- sum(forest.soil[,i])
}


soil<-summary.soil[with(summary.soil,order(-Count)),]
kable(soil, caption= "Soil Type Counts", format="pandoc", row.names = F)
```

```
st = stack(as.data.frame(forest.scaled[,forest.var.continuous]))
ggplot(as.data.frame(st)) +
  geom_boxplot(aes(x = ind, y = values)) +
  theme(axis.text.x = element_text(angle=45, hjust = 1)) +
  scale_x_discrete(name ="") + scale_y_continuous(name ="") +
  ggtitle("Boxplots of Scaled Continuous Variables")
```

```
density.plots = densityplot(~ Elevation + Aspect + Slope + HDist.Hydrology + VDist.Hydrology +
                              HDist.Roadway + Hillshade.9am + Hillshade.12pm + Hillshade.3pm +
                              HDist.FirePoint,
                            data=forest,
                            groups = CoverType,
                            plot.points = FALSE,
                            auto.key = list(space="right",title="Cover Type",cex=.6),
                            scales= list(x="free",y="free"),
                            xlab = '',
                            ylab=list(cex=.8),
                            aspect="fill",
                            par.strip.text=list(cex=.9))
plot(density.plots)
```

```
density.plots = densityplot(~ Elevation + Aspect + Slope + HDist.Hydrology + VDist.Hydrology +
                              HDist.Roadway + Hillshade.9am + Hillshade.12pm + Hillshade.3pm +
                              HDist.FirePoint,
                            data=forest.numeric,
                            groups = Z.Area,
                            plot.points = FALSE,
                            auto.key = list(space="right",title="Area",cex=.6),
                            scales= list(x="free",y="free"),
                            xlab = '',
                            ylab=list(cex=.8),
                            aspect="fill",
                            par.strip.text=list(cex=.9))
plot(density.plots)
```

```
corrplot(cor(forest[, forest.var.continuous]),
         tl.col = "black", tl.cex = 0.8, tl.srt = 45,
```

```
         cl.cex = 0.8, pch.cex = 0.8, diag = FALSE,
         type="lower",
         addCoefasPercent = TRUE, addCoef.col = TRUE,number.cex = .6) #Matt added to show correlation a
```

```
idx = grep("SoilType|CoverType", colnames(forest.numeric))
df = as.data.frame(forest.numeric[,idx])
idx.type = grep("CoverType", colnames(df))

df.temp = df[,-idx.type]

soil.sums = apply(df.temp,2,function(x) {
  tbl = table(x,df$CoverType)
  if (dim(tbl)[1] < 2) {
    tbl = rbind('0' = tbl, '1' = rep(0,7), deparse.level = 1)
  }
  return (apply(tbl,1,sum)[2])
})
#soil.sums

soil.sums.byclass = apply(df[,-idx.type],2,function(x) {
  tbl = table(x,df$CoverType)
  tbl = tbl[seq(2,14,by=2)]
  return (tbl)
})
#soil.sums.byclass

soil.ratios = as.data.frame(t(soil.sums.byclass)/soil.sums)
library(RColorBrewer)
soil.ratios.m = na.omit(as.matrix(soil.ratios))
barchart(soil.ratios.m,col=brewer.pal(7, "Pastel2"),xlab='',
         key=list(space="right",
                  lines=list(col=brewer.pal(7, "Pastel2"),lwd=3),
                  text=list(covertype.names)
))
```

```
idx = grep("Area1|Area2|Area3|Area4|CoverType", colnames(forest.numeric))
df = as.data.frame(forest.numeric[,idx])
idx.type = grep("CoverType", colnames(df))

df.temp = df[,-idx.type]

area.sums = apply(df.temp,2,function(x) {
  tbl = table(x,df$CoverType)
  if (dim(tbl)[1] < 2) {
    tbl = rbind('0' = tbl, '1' = rep(0,7), deparse.level = 1)
  }
  return (apply(tbl,1,sum)[2])
})
#area.sums

area.sums.byclass = apply(df[,-idx.type],2,function(x) {
  tbl = table(x,df$CoverType)
  tbl = tbl[seq(2,14,by=2)]
  return (tbl)
```

```
})
#area.sums.byclass

area.ratios = as.data.frame(t(area.sums.byclass)/area.sums)
library(RColorBrewer)
area.ratios.m = na.omit(as.matrix(area.ratios))
barchart(area.ratios.m,col=brewer.pal(7, "Pastel2"),xlab='',
         key=list(space="right",
                  lines=list(col=brewer.pal(7, "Pastel2"),lwd=3),
                  text=list(covertype.names)
                  )
)
```

```
ggplot(forest, aes(x=HDist.Roadway)) +
  geom_histogram(aes(group=CoverType, colour=CoverType, fill=CoverType), bins=30, alpha=0.7)+
  ggtitle('')+
  theme(legend.title = element_blank())+
  labs(x="Distance to Roadway",y="Count")
```

```
ggplot(forest, aes(x=HDist.Hydrology)) +
  geom_histogram(aes(group=CoverType, colour=CoverType, fill=CoverType), bins=30, alpha=0.7)+
  ggtitle('')+
  theme(legend.title = element_blank())+
  labs(x="Distance to Water",y="Count")
```