Team Checkpoint - Team 1

Stephan, David, Mike & Rahul

10/27/2018

Team Members

This report is for Team 1, which includes:

- Stephan Granitz
- Michael Kapelinski
- David Lyngholm
- Rahul Sangole

Introduction

The United States forest Service (UFS) considers itself a multi-faceted agency. Cumulatively, the agency manages and protects 154 national forests and 20 grasslands in 43 states and Puerto Rico including approximately 500 million acres of private, state and tribal forests, on which the UFS promotes sustainable management. The UFS mission is to sustain the health, diversity, and productivity of the nation's forests and grasslands to meet the needs of present and future generations. Some highlights of the UFS responsibility are shown below [1]:

- 13 billion dollars contributed to the U.S. economy by visitor spending each year
- 193 million acres managed by the Forest Service
- 27 million annual visits to ski areas on national forests
- 7.2 million acres of wetlands
- 36.6 million acres of wilderness
- 400,000 acres of lakes
- 57,000 miles of streams
- 10,000 professional wildland firefighters
- 154 national forests
- 20 percent of America's clean water supply provided by the national forests and grasslands

Like any managerial process, the UFS's success is measured in terms of how well it maximizes its use of resources (time, people, and money) to deliver value to their customers, in this case the US taxpayer/government. To maximize their impact, the UFS has had success creating partnerships with public and private agencies that help the UFS plant trees, improve trails, educate the public, and promote sustainable forest management and biodiversity conservation domestically and internationally.

Given the large area, various climates and terrains, diverse plants and forest species, and wildlife that make surveying the locations on any kind of frequent basis, the Forest Gladiator team has been tasked with developing a predictive model(s) using existing data to improve the UFS's management capabilities involving forest cover areas.

The Modeling Problem

There is a need to predict the forest cover for 30 x 30 meter cells using cartographic variables obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. The objective is to predict cover type as a multiclass classification problem based on the associated attributes (features).

The Data: Data Inventory and Data Quality Check

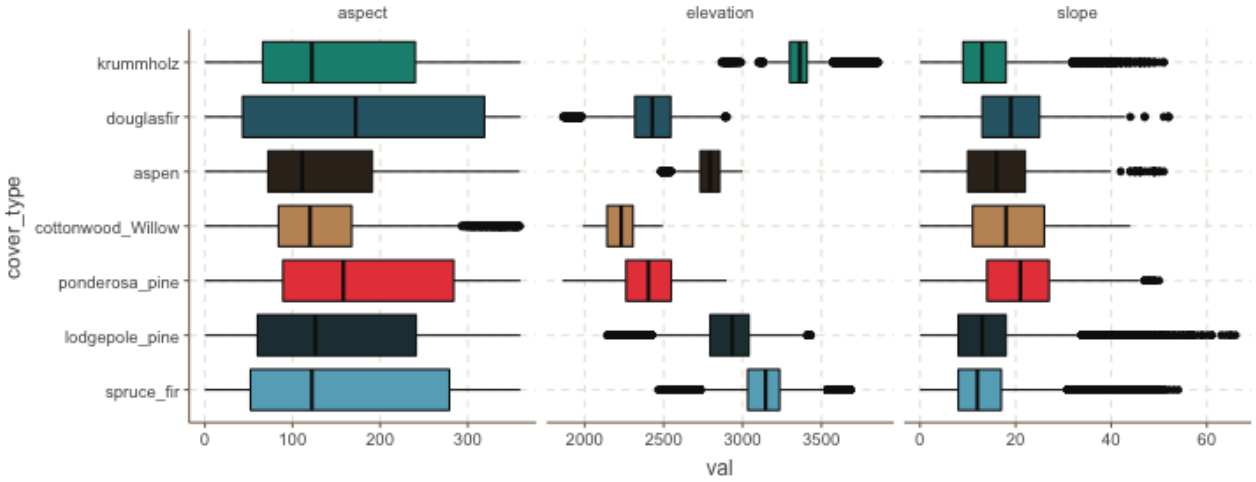
In our raw data it is clear that we don't have an even distribuion of the target variable (cover types). The data set is over 85% Spruce Fir and Lodgepole Pine. Another nearly 10% are Ponerosa Pine and Krummholz. While developing the training and testing data splits, this was taken into account. The splits were stratified on the cover types so that both data sets have equal distributions of cover type.

cover_type	n	percent
spruce_fir	211840	36.5%
lodgepole_pine	283301	48.8%
ponderosa_pine	35754	6.2%
cottonwood_Willow	2747	0.5%
aspen	9493	1.6%

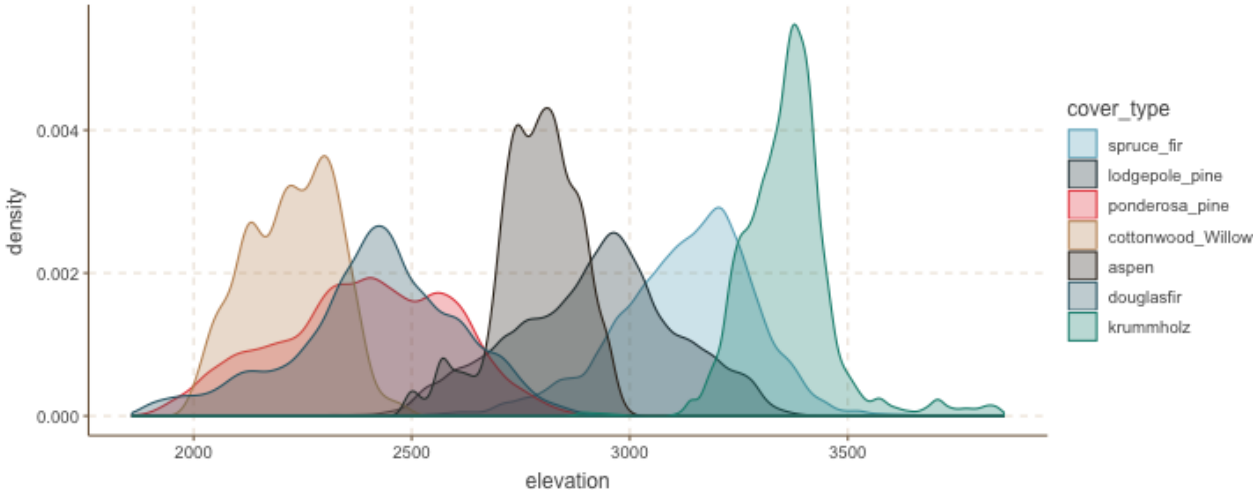
[1] Source: https://www.fs.fed.us/about-agency/newsroom/by-the-numbers

cover_type	n	percent
douglasfir	17367	3.0%
krummholz	20510	3.5%

From a quick skim of the data we see it is already pretty clean. There are no missing data and nothing jumps out as obviously out of the ordinary. There are clear groups of variables that we will explore for patterns such as soil type, distance measures, shade at different times of day, and wilderness area. If we set aside the larger groups of variables, we get three other variables to look at: aspect, elevation, and slope. A quick look shows that elevation may help separate the cover types. There may also be interesting ways to combine these variables.



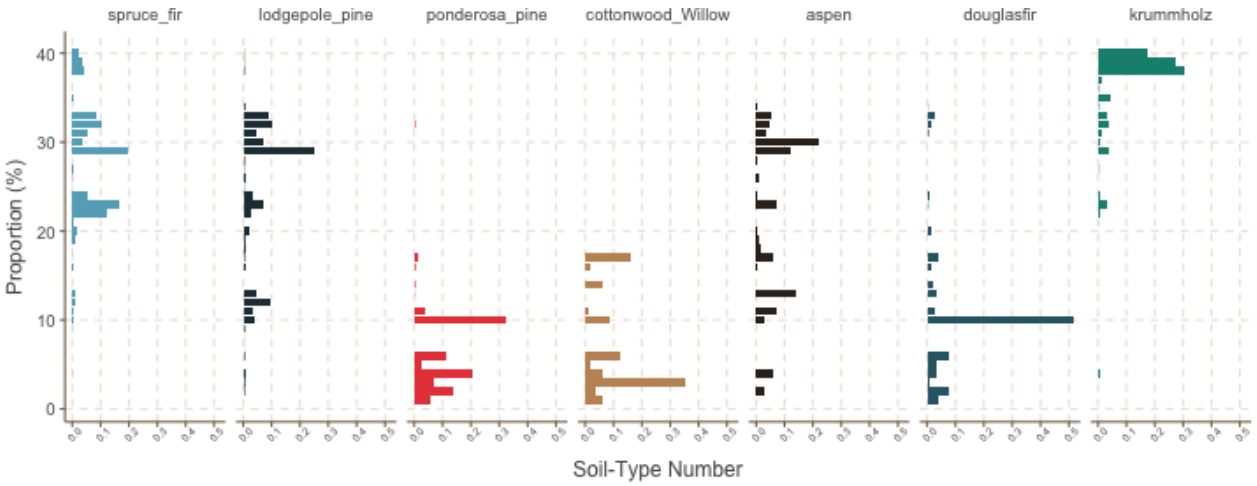
A density plot of elevation shows there is overlap but the groups are reasonably separated. This will likely be an important variable for our models.



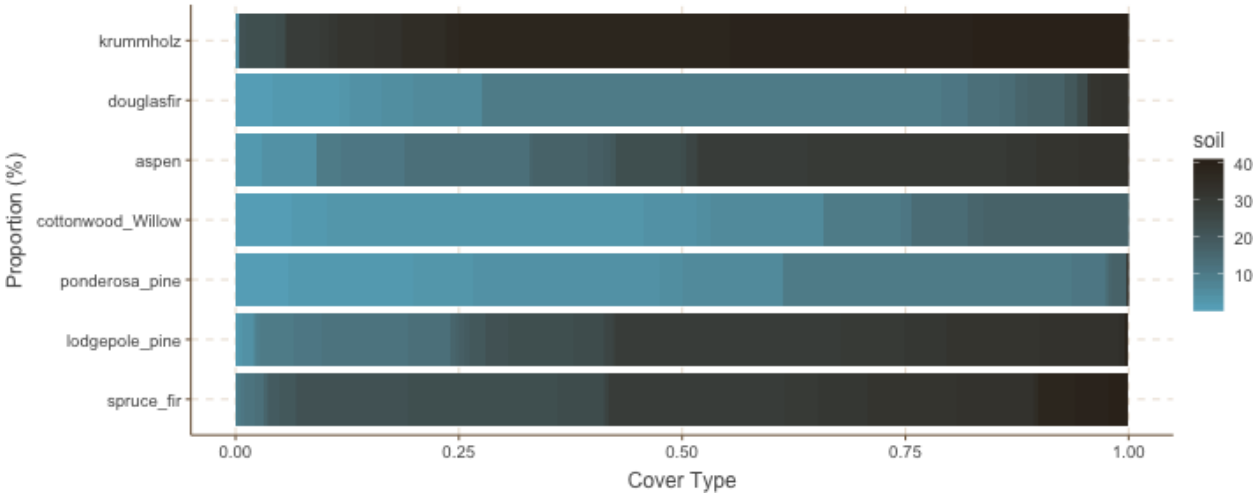
EDA: Initial exploratory data analysis.

The first group of variables are the 40 soil types. By plotting the proportion of each cover type contains each soil type we get some obvious clusters. Likely in modeling we will want to group the soil types. Initial insights show that Krumholz is mostly in soil types 30 through 40. Ponderosa Pine and Cottonwood Willow are 0 through 20.

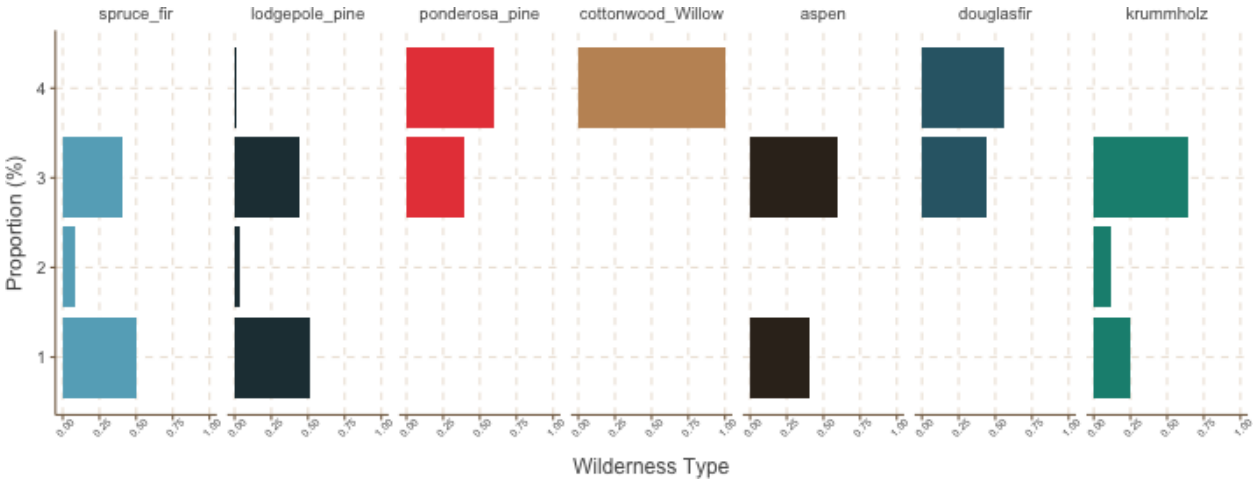
Soil types present in a cell can help eliminate the options for cover type in our model.



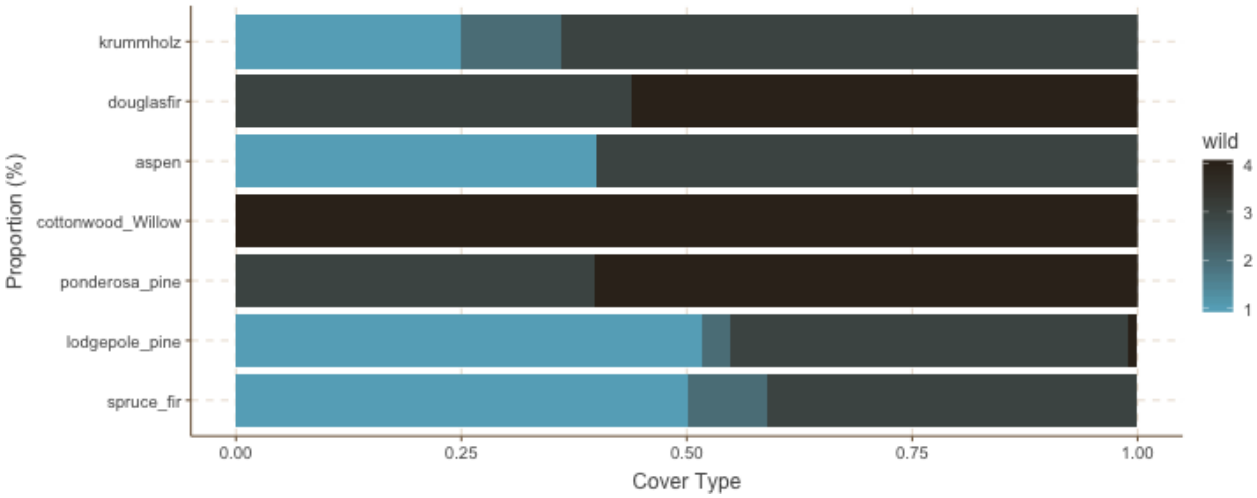
Looking further into the soil types we can see that cover types usually have mostly soil types over or under 20, not both, except for aspen which is spread across the spectrum.



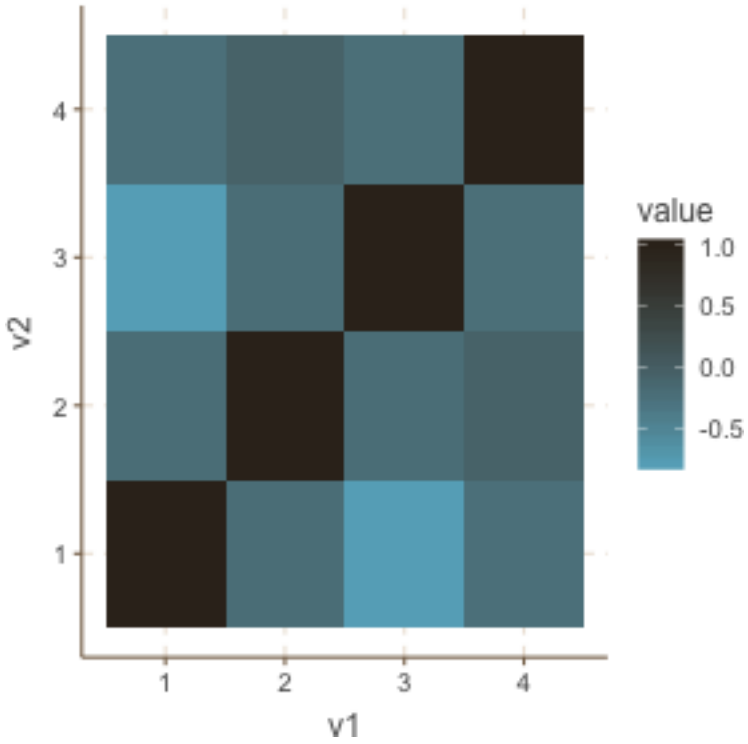
The next grouping of variables is wilderness area. Again we see some clear delineations aong the cover types. Cottonwood Willow is only in area four. Ponderosa Pine and Douglasfir have three and four, whereas the reaining cover types are primarily three and one.



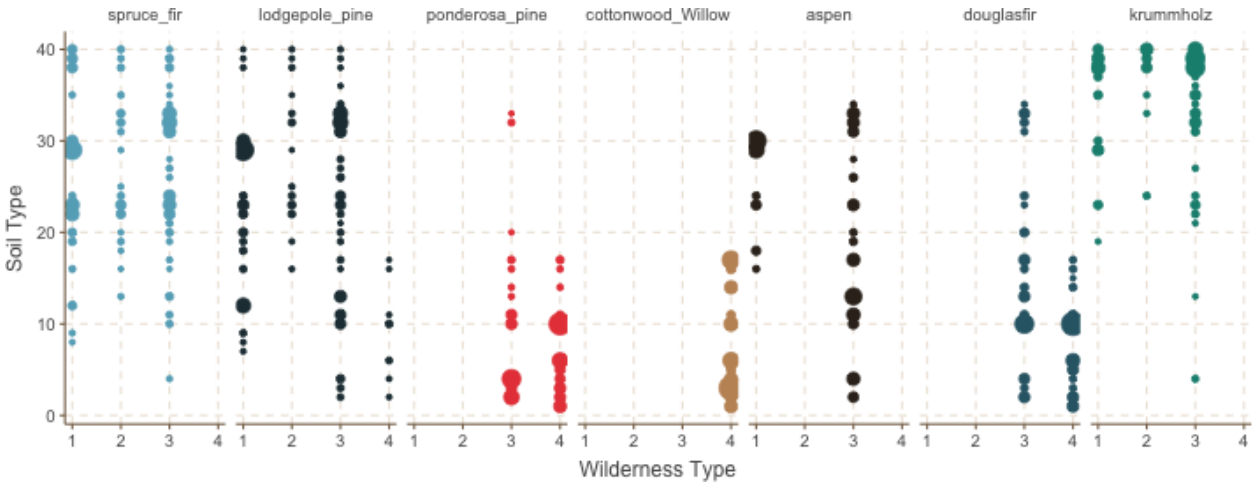
Another view of the proportions confirms this.



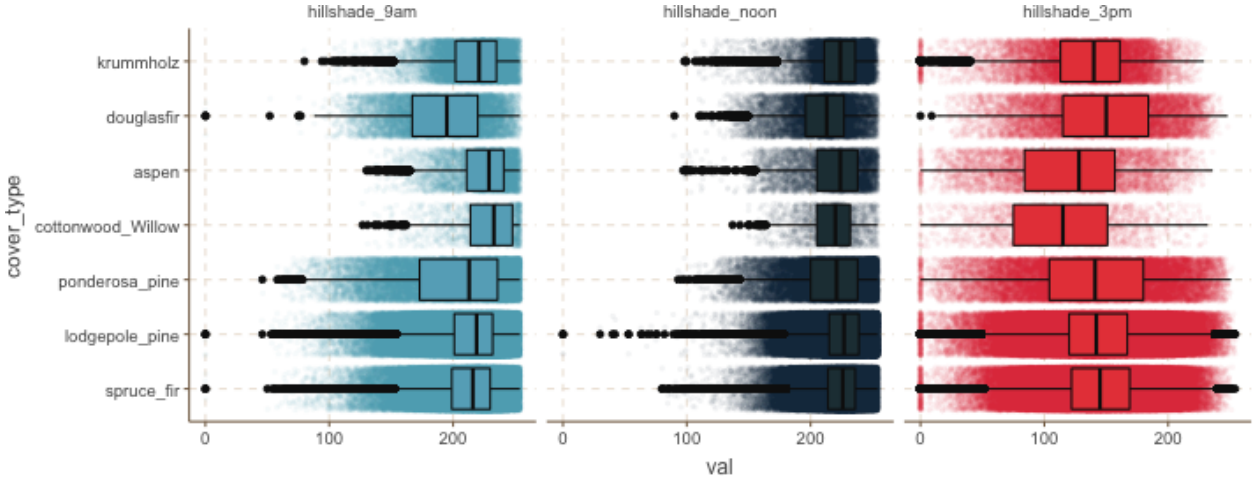
Interestingly wilderness areas three and one are negatively correlated. This may be useful for our models.



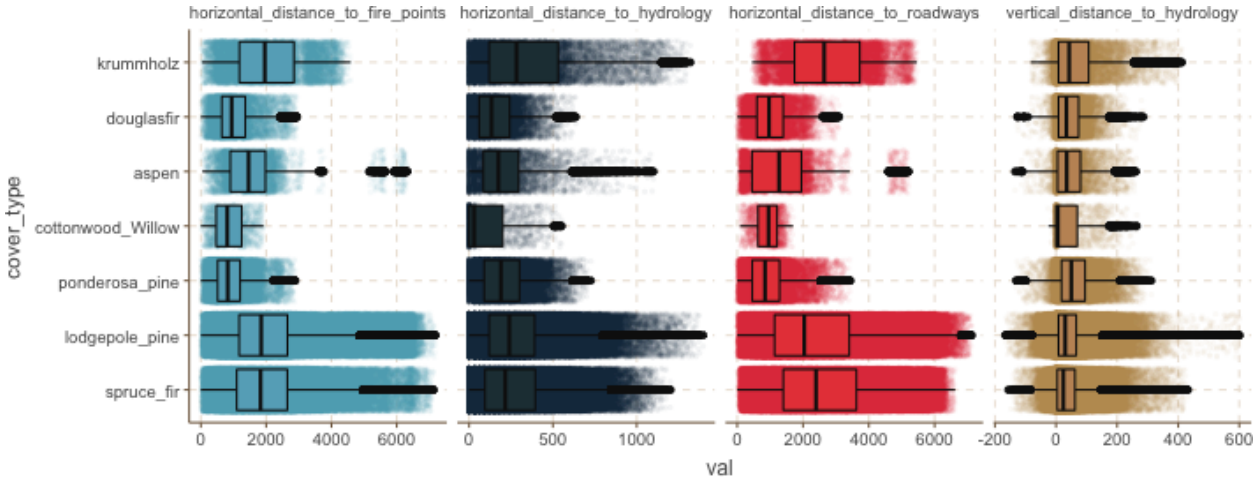
Combining soil types and wilderness areas into one plot gives us some obvious clusters in the training data. Further investigation of how to best group these variables will be useful for the modeling phase.



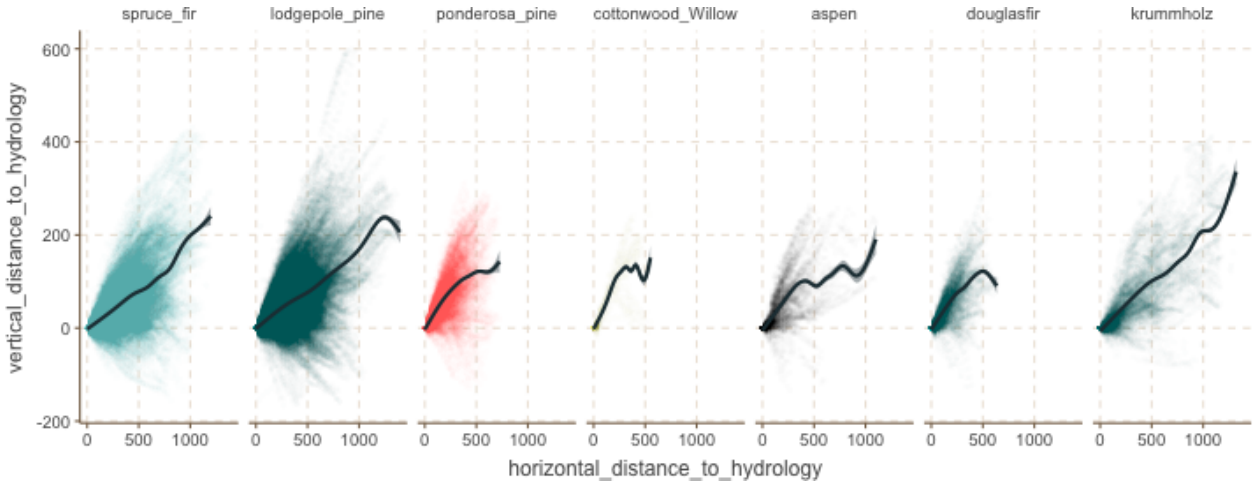
Shade measurements at various times of day show that some cover types have more variation than others. The pines and Spruce Fir are more likely to have lower shade values (<100) than the other cover types.



Similarly, the distance variables have cutoffs that can help identify a cover type. Lodgepole Pine and Spruce Fir have much wider ranges of distance measurements than the other cover types. There are also quite clear upper limits for different types of cover types. This is leveraged to create new feature variables. For example, only Lodgepole has vertical distances to hydrology > 450.



Looking at the horizontal and vertical distance to hydrology, most cover types have similar, positive correlations. Combining these variables may further help in distinguishing our target variable. Some cover types have large values on both the positive and negative side, while others have tighter distributions. This is also exploited in feature creation.



Feature Engineering and Data Preparation

Based on the EDA performed, many new features are appended to the raw data:

1. Transformation of continuous variables to binned variables is performed since some tree based learners do better with binned variables.
2. Wilderness areas and Soil Types are also added as factor variables, with 4 and 40 levels respectively.
3. Distance based features are added, viz:
   - Straight line distance (eucledian) : Square root of sum of squares of `horizontal_distance_to_hydrology` and `vertical_distance_to_hydrology`
   - Indicator variable for if `vertical_distance_to_hydrology` < 0
   - Indicator variable for if `vertical_distance_to_hydrology` $<= 350$
   - Indicator variable for if `vertical_distance_to_hydrology` is between 350 and 500
   - Indicator variable for if `vertical_distance_to_hydrology` $>= 500$
   - Indicator variable for if `horizontal_distance_to_hydrology` < 600
   - Indicator variable for if `horizontal_distance_to_hydrology` < 1250
   - Squared Distance Ratio : equal to the square of the ratio of `horizontal_distance_to_hydrology` and `vertical_distance_to_hydrology`
4. Principal Components:
   - 6 PCA components for all variables except the Soil and Wilderness variables
   - 2 PCA components for the Hillshade variables

A few data transformations are also performed:

1. BoxCox transformations of continous variables is investigated to make their distributions normally distributed. `Elevation` is found to have a lambda value of 2.
2. Standardized (centered and scaled) continuous variables

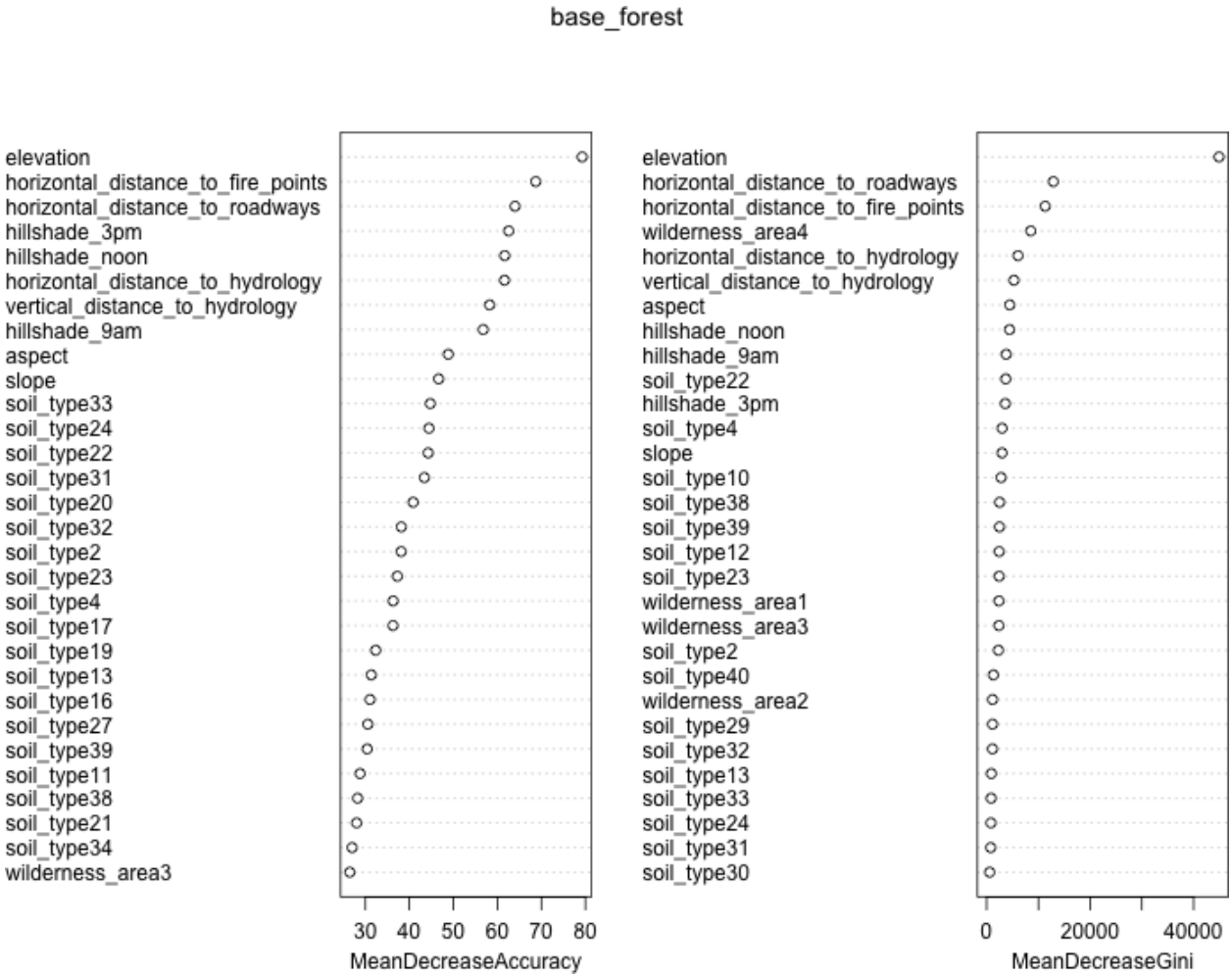
Certainly, not all of these features & transformations are applied at the same time. The creation of the features / transformations is modularized using functions. As a result, various configurations of the data can be easily investigated.

Predictive Modeling: Initial modeling results

A preliminary Random Forest model is constructed on the original raw dataset (without much pre-processing or feature engineering) to gain a some understanding of the key relationships between response and predictors, as well as establish a baseline performance one can expect. The performance of the model is decent, with a 16% error rate for the OOB samples. Class errors vary, from 8% for Ponderosa Pune to 80% for Aspen. However, these are both the minority classes which indicates some careful work might be needed to get adequate performance for the smaller classes. For the two majority classes, Spruce Fir and Lodgepole Pine, the performance is ~15%.

```
##
## Call:
##  randomForest(formula = cover_type ~ ., data = cover_data, mtry = sqrt(ncol(cover_data)),      ntree
##                Type of random forest: classification
##                      Number of trees: 300
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 16.12%
## Confusion matrix:
##                   spruce_fir lodgepole_pine ponderosa_pine
## spruce_fir            122349          25223             63
## lodgepole_pine         18407         177880           1714
## ponderosa_pine             2           1774          22919
## cottonwood_Willow          0              0            831
## aspen                    332           4642            273
## douglasfir                14           2276           5149
## krummholz               3399             83              2
##                   cottonwood_Willow aspen douglasfir krummholz class.error
## spruce_fir                        0     2         10       641  0.17492312
## lodgepole_pine                    5    79        151        75  0.10302505
## ponderosa_pine                   64     3        266         0  0.08426562
## cottonwood_Willow              1085     0          7         0  0.43577743
## aspen                             0  1388         11         0  0.79115257
## douglasfir                       55     0       4663         0  0.61643498
## krummholz                         0     0          0     10873  0.24266908
```

The variable importnace plot builds upon what is seen in the EDA, and indicates what features should be pursued further. The plot on the left indicates which variables' inclusion reduces the classification error rate. The greater the MDA (Mean Decrease in Accuracy) the more important this variable is. The plot on the right indicates which variables contribute to the highest homogeneity in the nodes. The greater the mean decrease in Gini the higher the variables contribution to node purity. In both, we can see that `elevation` is clearly the strongest predictor, followed by some of the `distance` variables.



Next Steps: List of Modeling Approaches

The team has decided to pursue four types of models going forward:

1. Random Forest
2. Boosting using `xgboost`
3. Support Vector Machines
4. Bayesian Modeling

A few preliminary RF models are also going to be used to sort through the newly created features in an attempt at feature reduction and exploratory analysis into the data. Thereafter each individual model will be built on a number of types dataset preparations (for example: using dummy variables for factors vs using factors themselves, etc). The team will use techniques like repeated 10-fold Cross Validation to get estimates of test data performance, while using a metric like log-loss to evaluate the performance of any give model. Cross validation will also enable estimation of the standard error of the metric, as well as aid hyperparameter tuning.

If the team has an opportunity, we would also like to explore ensemble models which combine the output of each of the four models in an attempt to improve the predictive performance.