Team Checkpoint 3

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 multiclass 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

Most classification algorithms can be applied, either directly or through slight modifications, to multiclass classification problems. Two different approaches exists for multi-class classification: algorithms that naturally permit the use of more than two classes and those that first reduce a multi-class problem to a collection of binary-class problems and then combine their predictions in various ways. The following is a list of algorithms we will consider for our problem:

- Discriminant Analysis
- Logistic Regression
- K-Nearest Neighbors
- Random Forests
- Neural Network
- Support Vector Machine

We will use 'Kappa' as a metric to optimize for tuning our parameters because this metric can improve the quality of the model for problems where there are a low percentage of samples in one class.

2.1 Evaluating Classification Models

We will evaluate a number of different models by applying them to a holdout 'test' dataset and assessing their performance. Because there is an imbalance in the classes to be predicted, we will consider alternatives to 'accuracy' to evaluate our final models. Using accuracy only to evaluate models can perform poorly in predicting the less frequent classes. We will instead look at two metrics that handle class imbalance: balanced accuracy and f1 score. Balanced accuracy is (sensitivity+specificity)/2, or the average accuracy over all

classes. F1 Score is precision x recall/(precision + recall),the weighted average of precision and recall. The precision measures the accuracy of a predicted positive outcome and recall (sensitivity) measures the strength of the model to predict a positive outcome Therefore, this score takes both false positives and false negatives into account.

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
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.

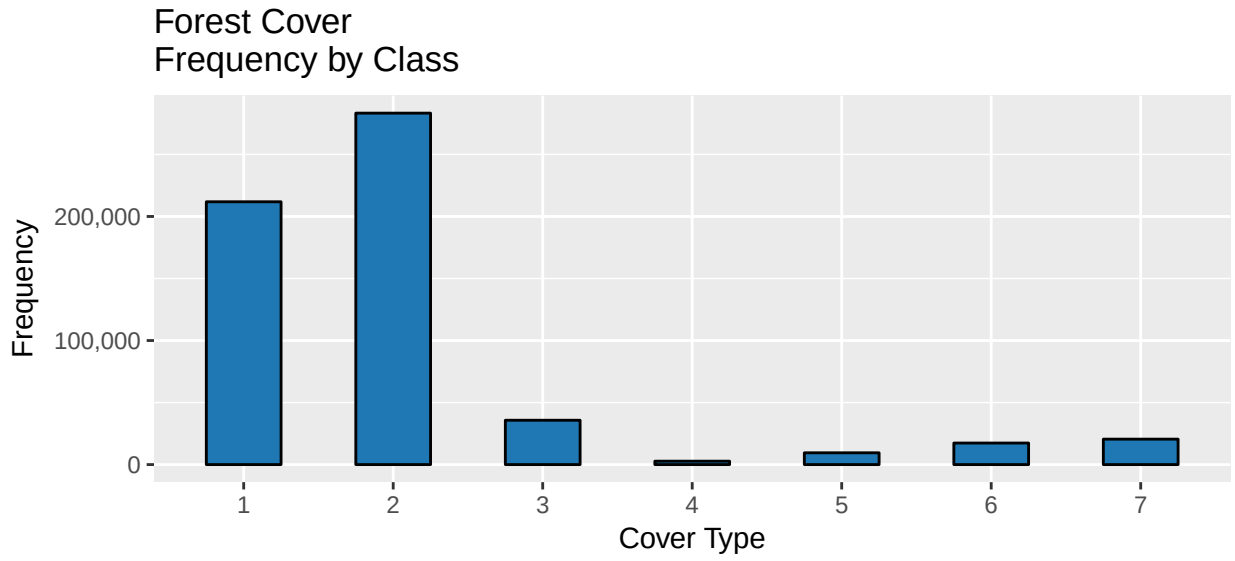


Figure 1: Frequency of each Cover Type Class

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	5260796
Area3	253364
Area4	36968
Area2	29884

Name	Count
SoilType13	17431
SoilType38	15573
SoilType39	13806
SoilType11	12410
SoilType4	12396
SoilType20	9259
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.

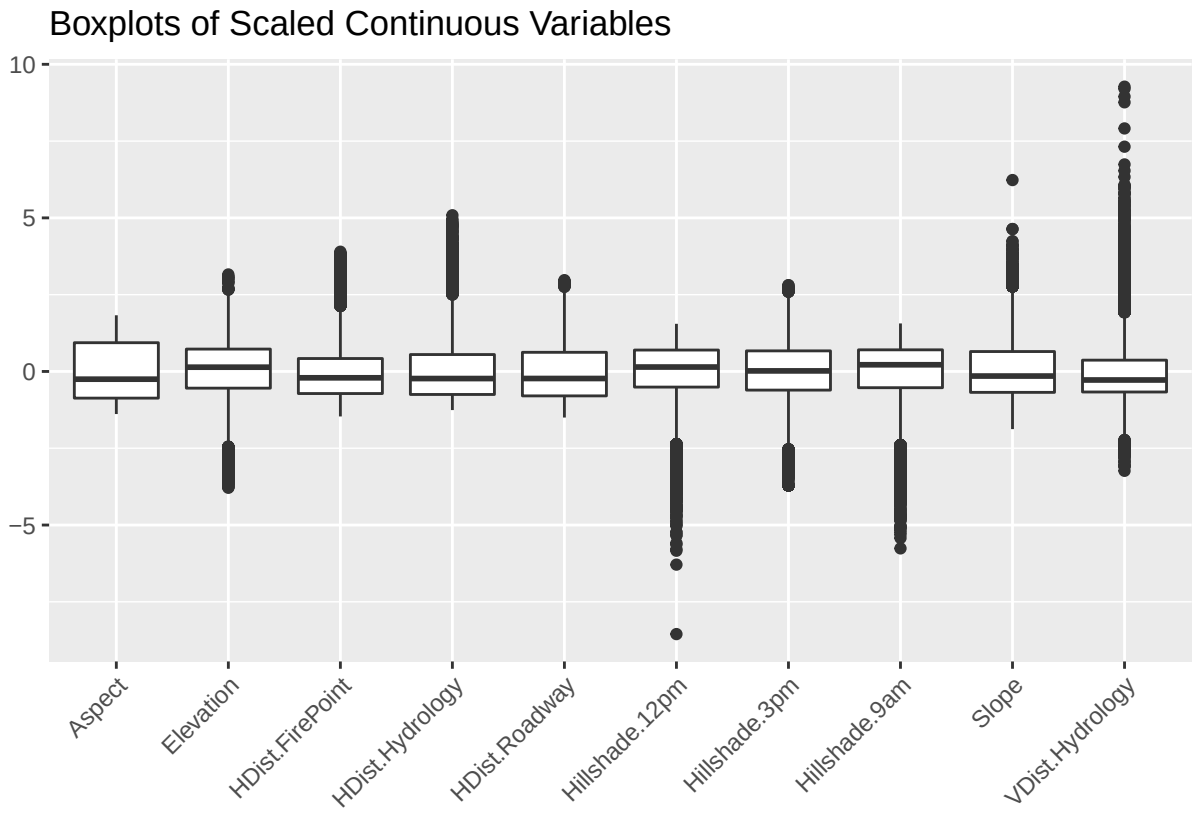


Figure 2: Boxplots of Scaled Continuous Variables

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.

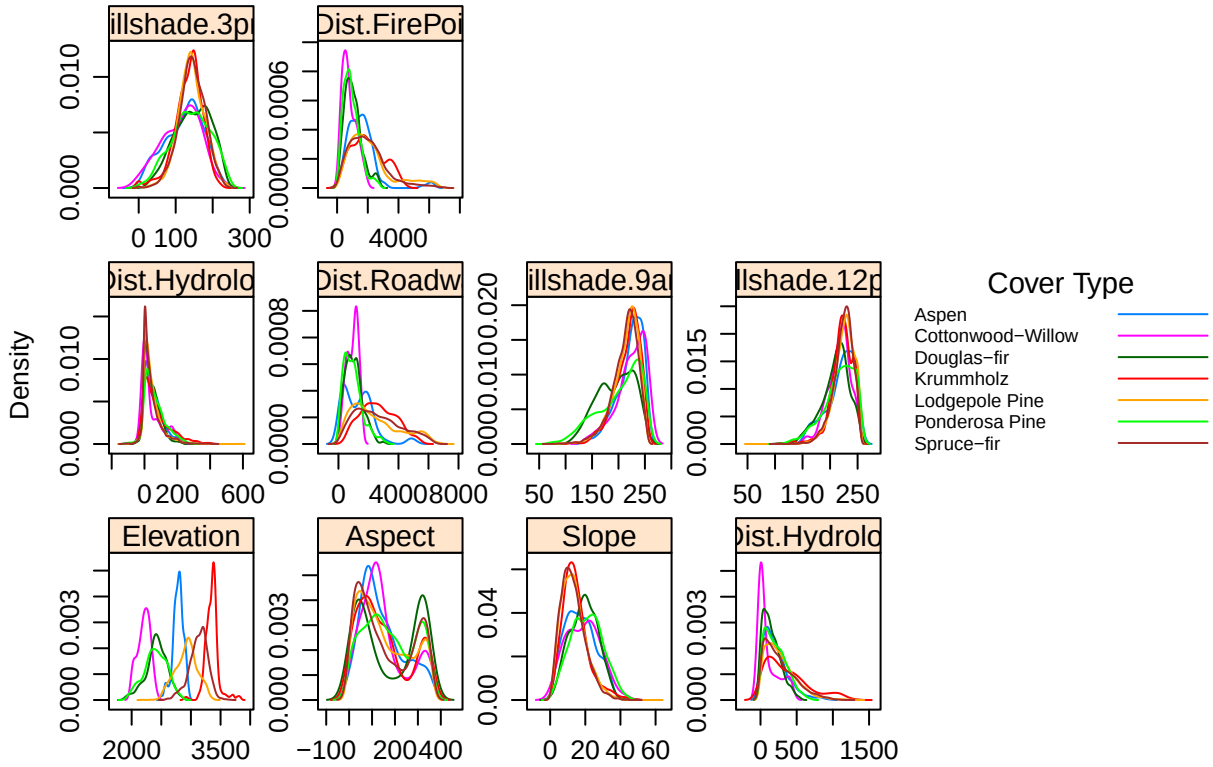


Figure 3: Density Plots of Continuous Variables by Class

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

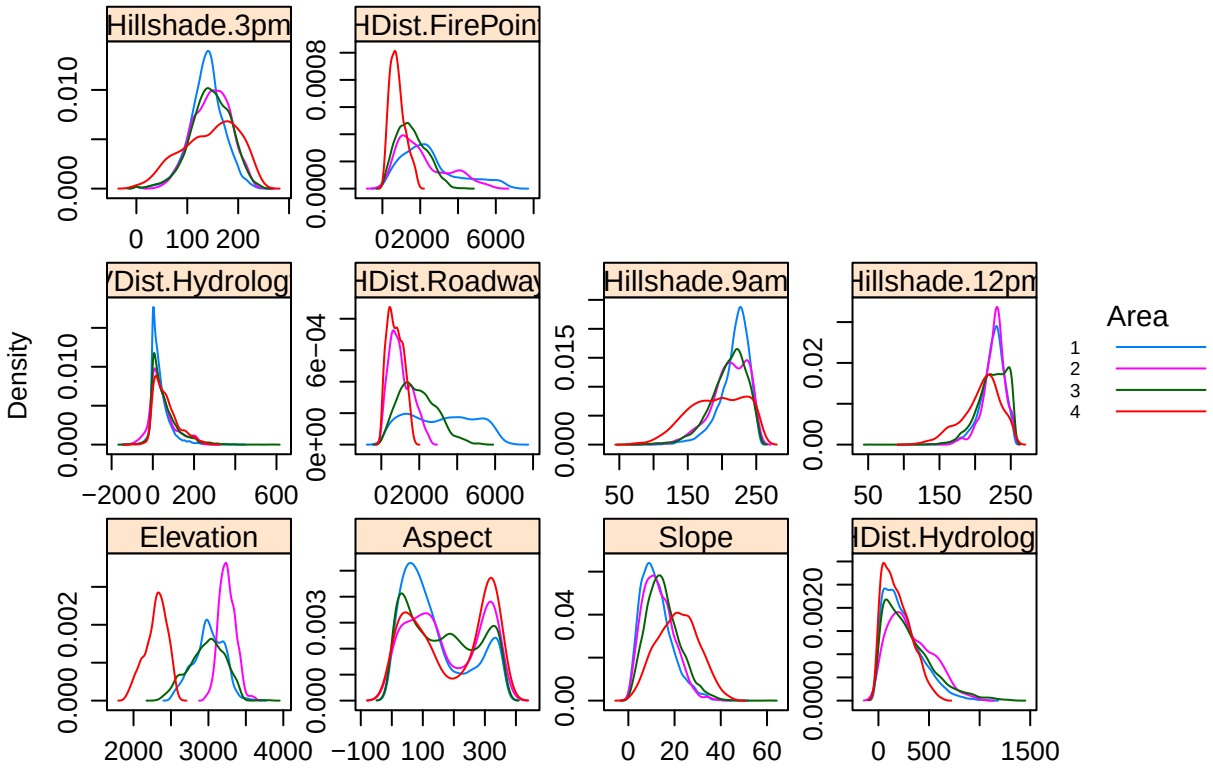


Figure 4: Density Plots of Continuous Variables by Area

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

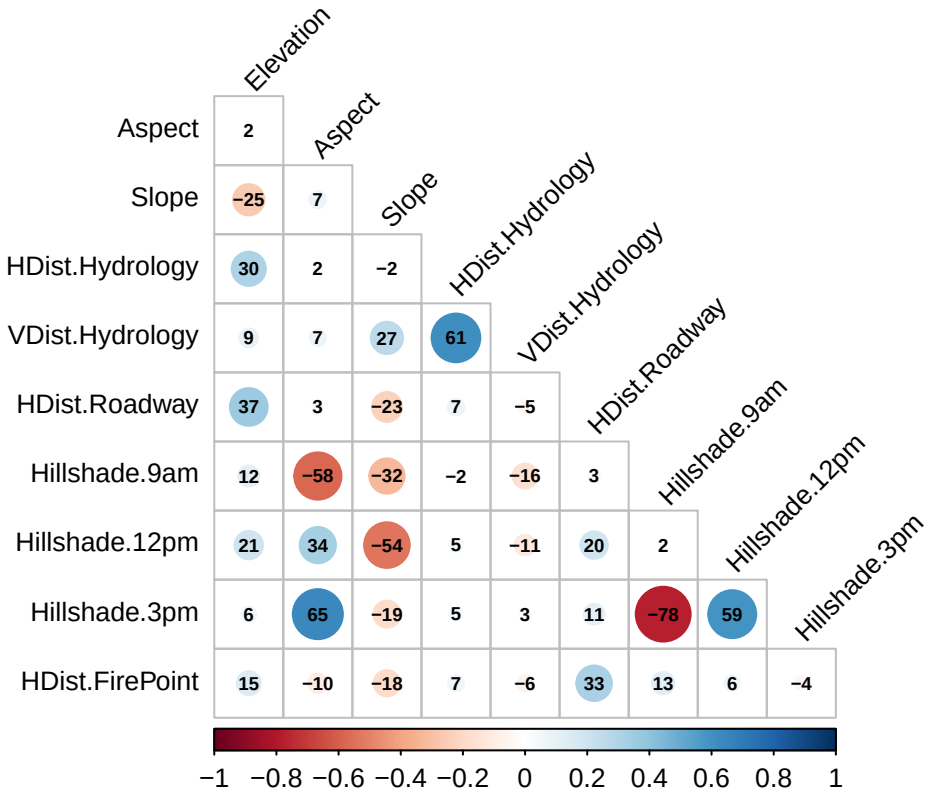


Figure 5: Correlation Matrix

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.

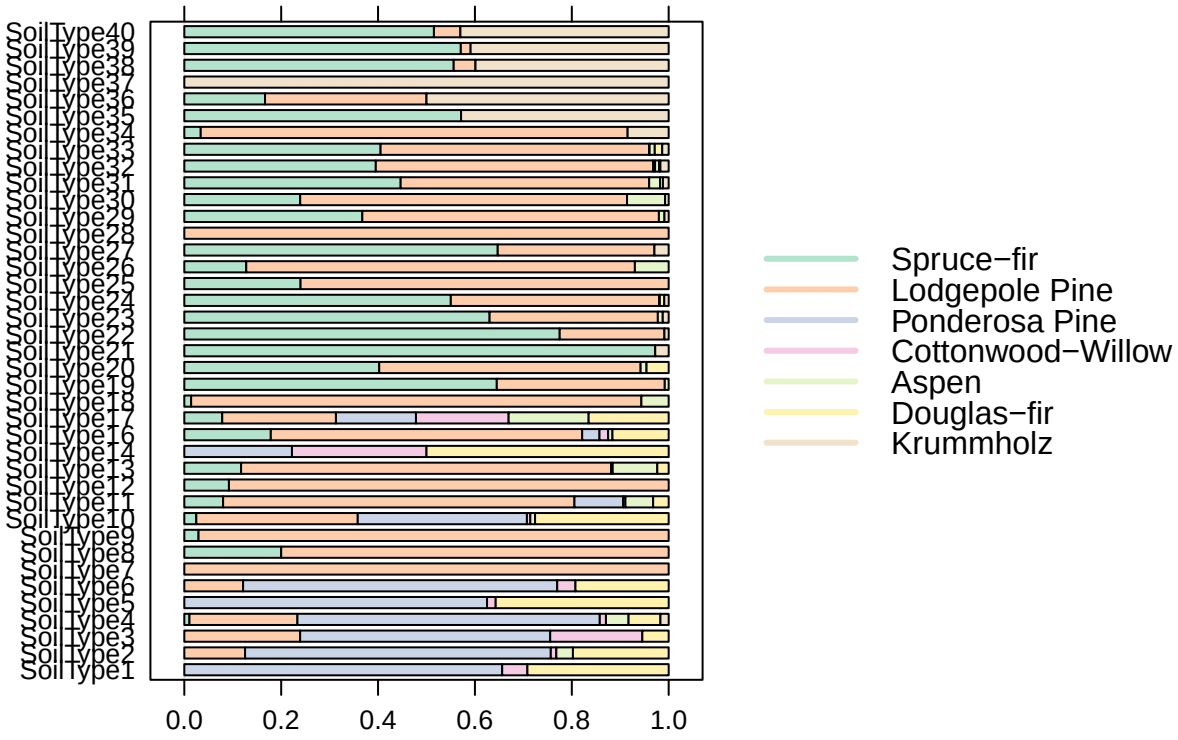


Figure 6: Soil Barchart

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.

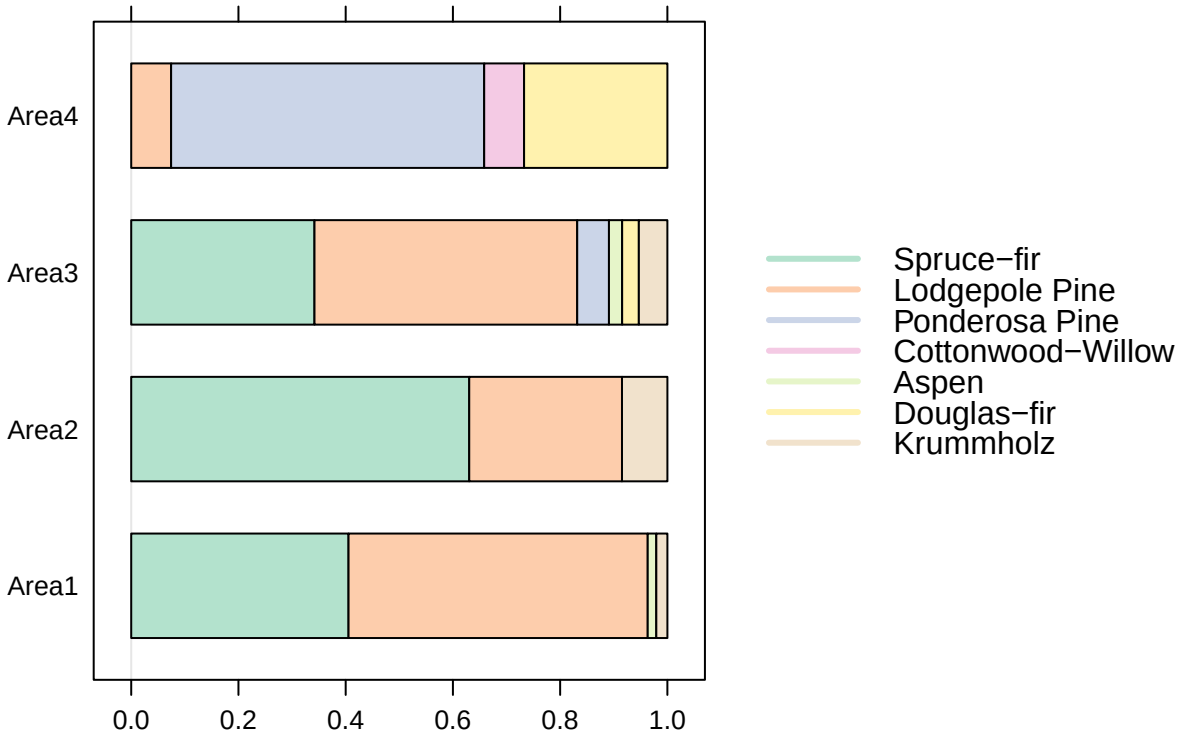


Figure 7: Area Barchart

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.

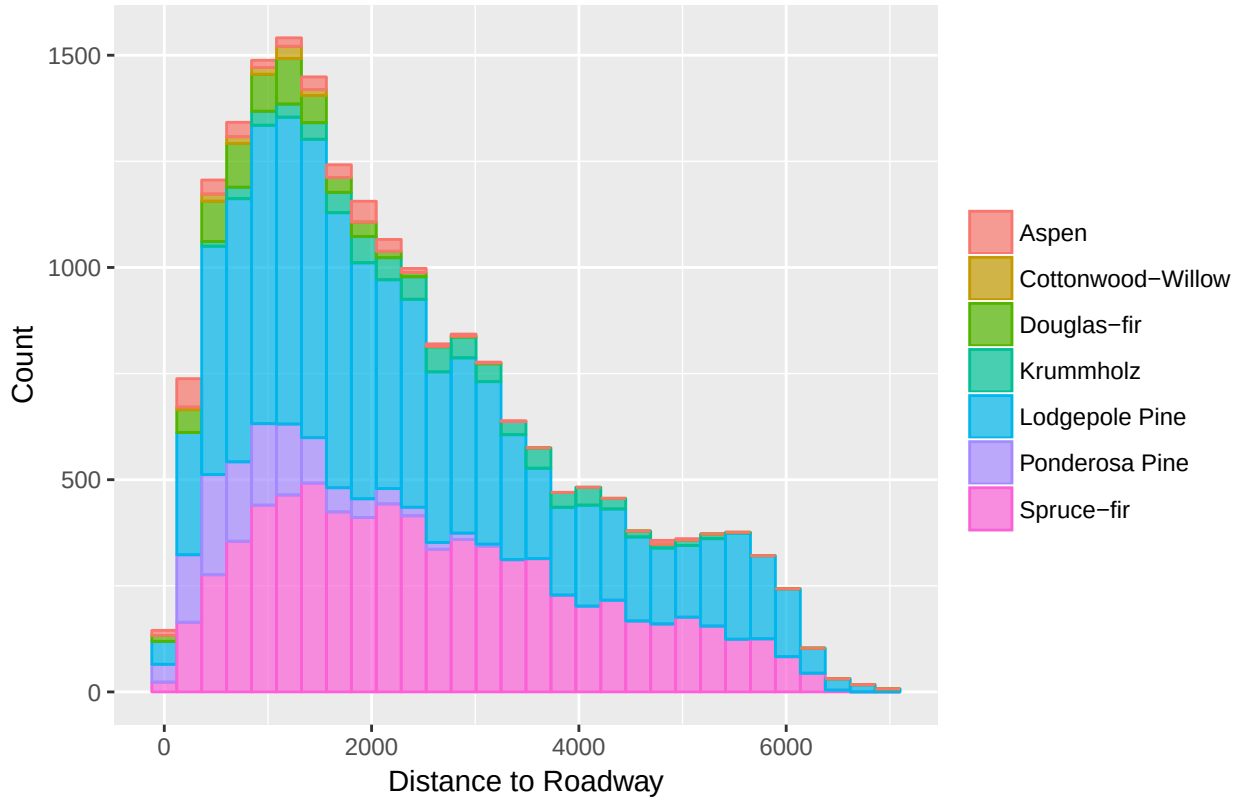


Figure 8: Distance to Roadway, by class

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.

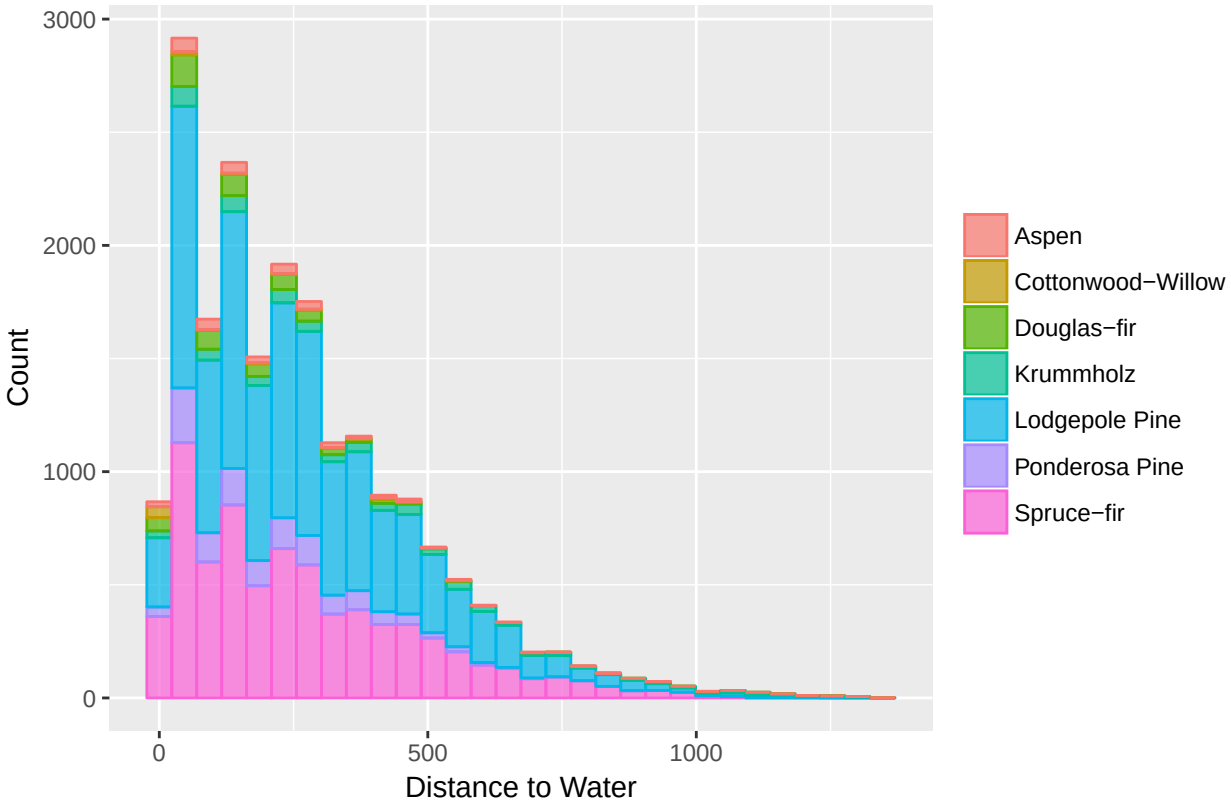


Figure 9: Distance to water, by class

Figure 10 shows the contrasting relationship between Hillshade.9am and Hillshade.3pm over varying aspect values. We see that between 0 and 200 degrees, hillshade varies significantly between 9am and 3pm, but between 200 and 360 degrees, hillshade between 9am and 3pm are relatively the same.

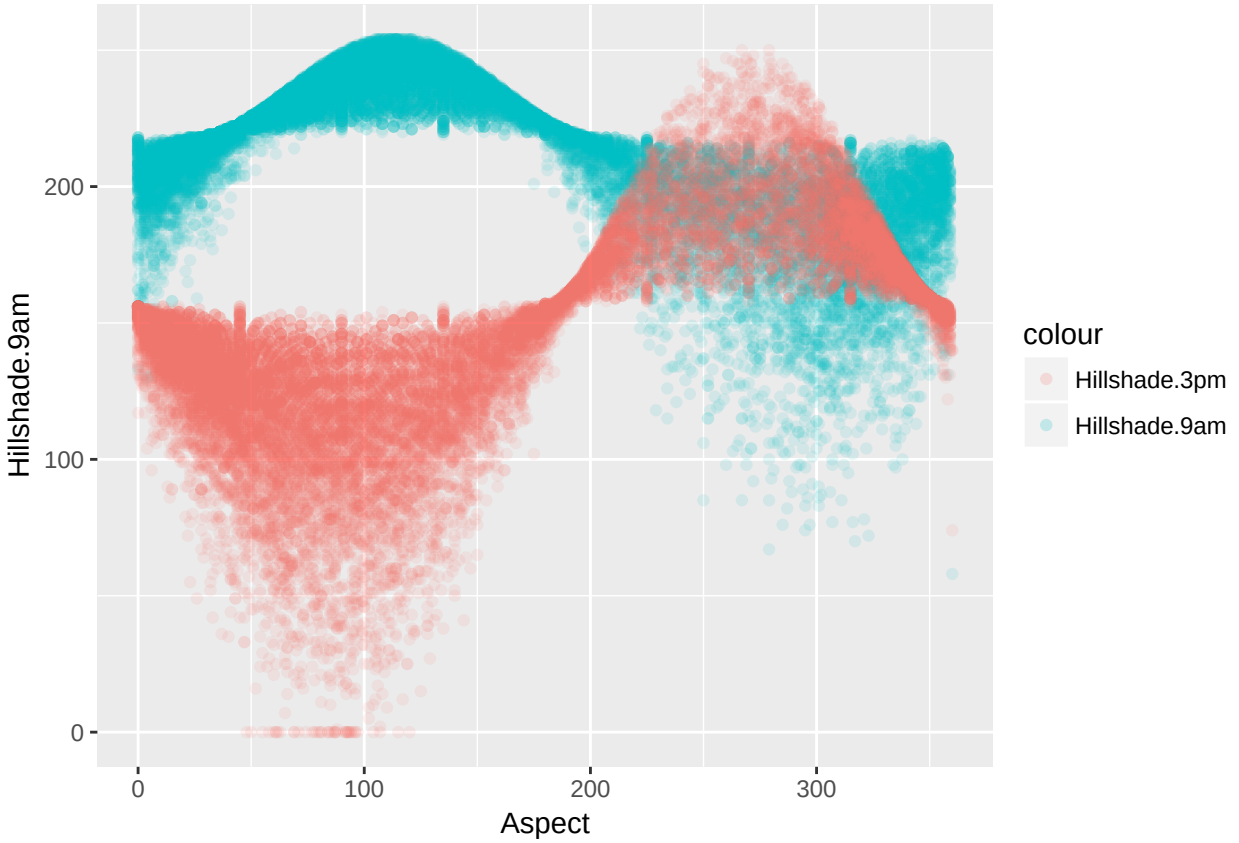


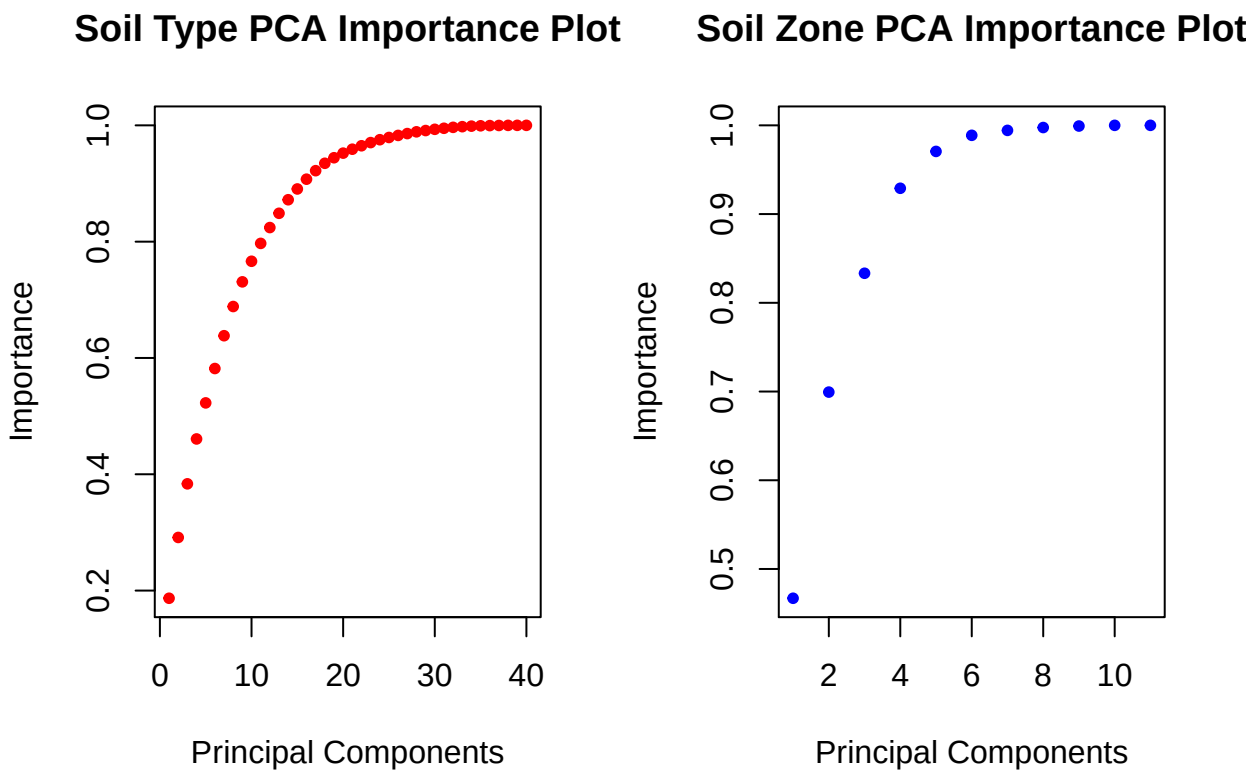
Figure 10: Hillshade.9am and Hillshade.3pm

5 Data Preparation

5.1 Transformations

The distribution for Hillshade.12pm was noticeably skewed in the boxplots above, so we create a log transformation of Hillshade.12pm and for later use in our models. Additionally we created an interaction variable using Vist.Hydrology and HDistHydrology. A linear distance variable was also created using the vertical and horizontal distance to hydrology variables.Additionally, soil types were grouped into Climatic and Geologic zones as a potential means to reduce the amount of SoilType variables used for various modeling techniques.

Furthermore, we attempt to reduce the variable complexity of 40 soil variables with principal component analysis. We found that 90.7% of the variance can be explained with 16 variables instead of 40. Similar analysis using 6 of 10 Soil Zones explains 99.2% of variation. This may help with modeling efforts the rely on reducing model complexity.



The correlation plot below also shows high correlations between Hillshade, Aspect, and Slope. By performing PCA analysis we are able to reduce 5 variables down to 3, with 0 correlation and explaining 99.7% of the variance. This is highly likely to be useful for models that need to minimize highly correlated variables.

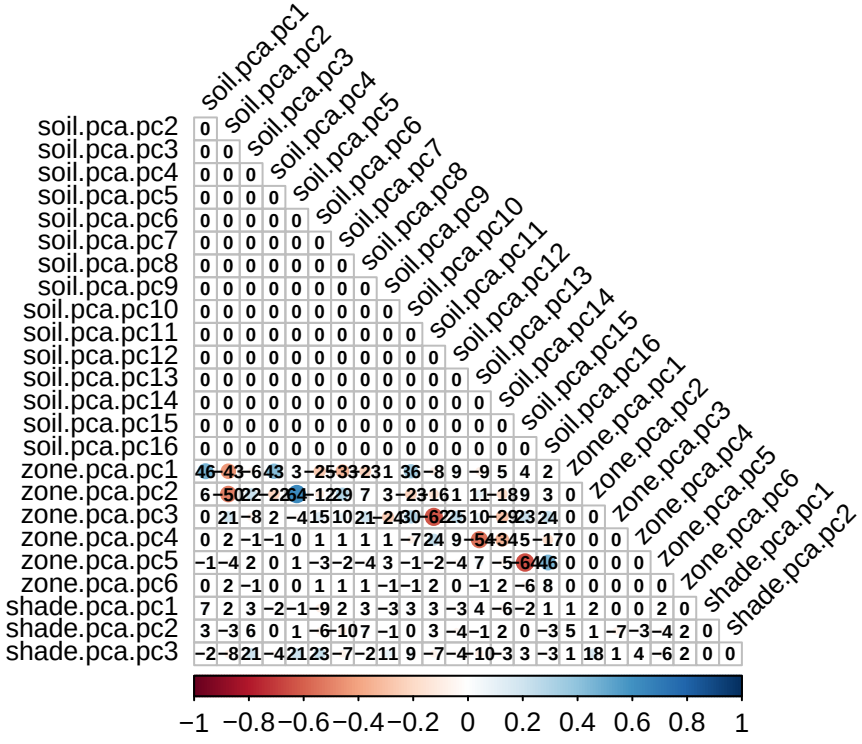


Figure 11: PCA Correlation

Several of the multi-classification models we will rely on accounting for variables with near zero variance. Neural networks, logistic regression and KNN models prefer that no near zero variance variables are in the dataset. In this dataset, 32 of the 40 soil variables have near zero variance. We will look to principal component analysis, combining predictors, and removing predictors to remedy this issue.

5.2 Splitting the Data

We have split our data into 70/30 training-testing set so that we could evaluate the performance of our models.

6 Exploratory Models - Feature Selection

Next we will run a few exploratory models to help us understand the relationships in our data and help to evaluate the most important variables for feature selection.

6.1 Linear Discriminant Analysis

Linear Discriminant Analysis tries to find a linear combination of the predictors that gives maximum separation between the centers of the data while at the same time minimizing the variation within each group of data. We ran an exploratory model on the data that returned the linear combinations of the original variables that were created to distinguish between the classes. The variables with large absolute values in the scalings are more likely to be influential. For this data, Elevation, Area4 and trans.zone27 seem to be among the important variables. The first discriminant function, LD1, achieves 73.85% of the separation of classes, with the second discriminant function, LD2, improving the separation by 18.19%. Therefore, to achieve a good separation of the classes, we should use both of the first two discriminant functions.

	LD1	LD2	LD3	LD4	LD5	LD6
Elevation	-0.988	0.869	-1.156	-0.081	-0.104	-0.088
VDist.Hydrology	0.026	-0.021	0.110	-0.102	0.227	-0.201
HDist.Roadway	0.127	-0.017	0.083	0.072	-0.090	0.471
HDist.FirePoint	-0.100	0.055	0.128	0.015	0.002	0.347
Area2	0.141	-0.019	0.081	0.070	-0.081	0.154
Area3	0.124	0.256	0.230	0.181	-0.148	0.002
Area4	0.721	0.787	-0.534	0.642	-0.017	0.574
trans.LDist.Hydrology	0.108	-0.183	0.159	-0.108	-0.068	0.325
trans.zone27	0.331	0.174	-0.178	-0.461	0.043	-0.021
trans.zone35	-0.007	-0.019	0.010	-0.004	-0.005	0.015
trans.zone42	-0.062	0.009	-0.014	-0.007	-0.032	0.133
trans.zone47	0.032	-0.066	0.041	0.007	-0.217	-0.055
trans.zone51	0.094	0.057	-0.074	0.432	0.211	0.131
trans.zone61	0.047	0.015	-0.034	0.162	0.187	-0.168
trans.zone67	-0.061	-0.008	-0.004	-0.016	0.044	-0.114
trans.zone71	-0.033	-0.008	-0.061	0.029	-0.026	-0.027
trans.zone72	-0.055	0.009	-0.168	0.020	0.002	-0.069
trans.zone77	-0.074	-0.173	0.027	0.053	0.092	0.133
trans.zone87	-0.016	0.405	0.432	-0.003	0.007	0.022
zone.pca.pc1	0.057	0.134	-0.039	-0.044	-0.113	-0.126
zone.pca.pc2	-0.059	0.054	-0.146	0.008	0.148	-0.013
zone.pca.pc3	0.124	0.340	0.223	-0.155	0.111	0.024
zone.pca.pc4	0.268	-0.125	-0.420	-0.347	0.047	-0.031

	LD1	LD2	LD3	LD4	LD5	LD6
zone.pca.pc5	0.058	0.055	0.083	-0.107	-0.015	0.042
zone.pca.pc6	0.034	-0.003	-0.032	0.184	0.199	-0.151
shade.pca.pc1	0.017	-0.062	-0.014	-0.073	-0.253	0.000
shade.pca.pc2	-0.038	-0.082	-0.065	-0.033	-0.422	0.277
shade.pca.pc3	0.013	-0.152	0.331	0.010	0.543	0.359

6.2 Decision Tree

The tree plot below shows the result of fitting a decision tree algorithm to all of our data, the purpose of which is to draw insights regarding predictor variables that could be most effective in building a predictive model. Each node in the tree shows the predicted class, the predicted probability of each class and the percentage of observations in the node. Our tree uses the variables Elevation, Trans.zone87 and Area3 to predict our classes.

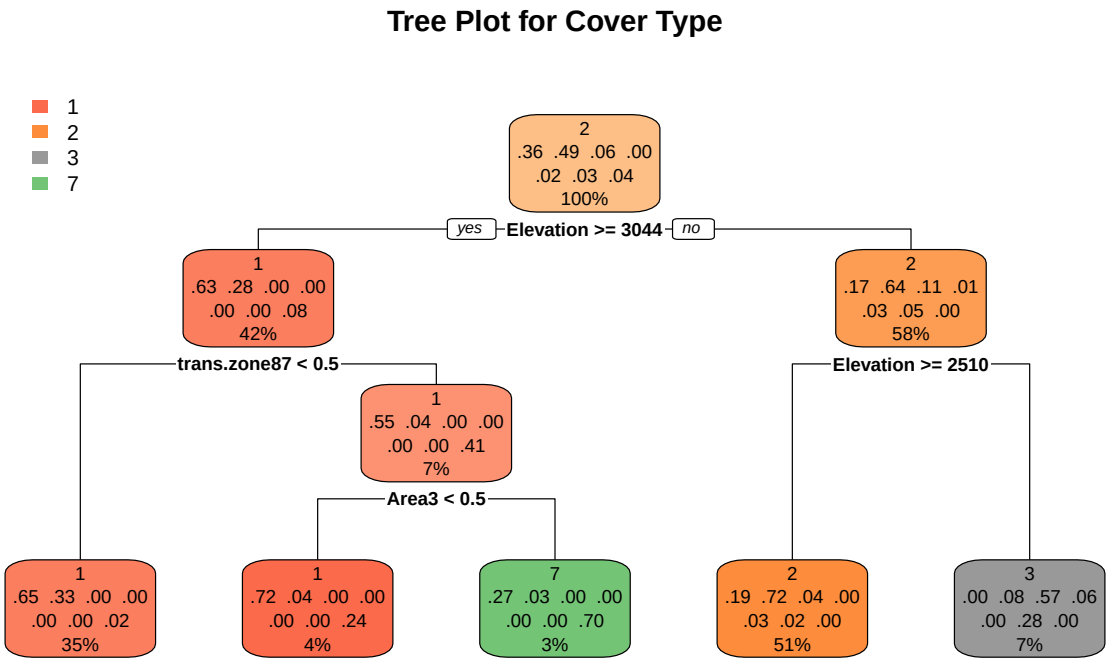


Figure 12: Tree Plot

6.3 Random Forest

The chart below shows the variables plotted by two measures of importance, Mean Decrease Accuracy and Mean Decrease Gini. Gini importance measures the average gain of purity by splits of a given variable.

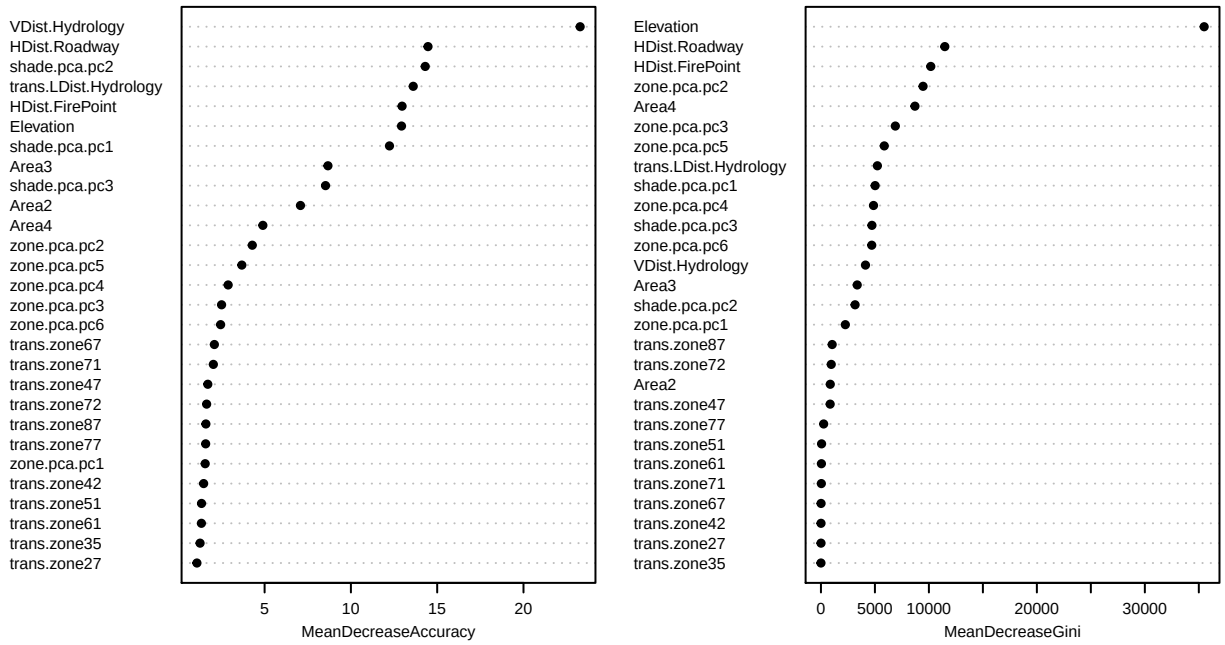


Figure 13: Variable Importance

7 Modeling Plan

Based on our modeling problem and exploratory data analysis, we will run various classification algorithms that can predict more than two classes. We will try variations of the following algorithms, which will require different data preparations.

7.1 Samples

In order to perform an initial analysis of the various models, we will use a sample of the training data. Once we have honed in on the correct tuning parameters for the models, we will run the models against the full dataset. This is an important step towards making the modeling process perform faster.

To accomplish this, we attempted the following samples:

1. 1900 observations from each class
2. random sample of 13,000
3. 5000 observations from each class with replacement
4. a tiered selection of observations, using all observations from low volume types, and using a higher sampling from the more popular types
5. a random sampling of 40,000
6. Stratified sample of 23,000
7. Stratified sample of 46,000

7.2 Neural Network

Rather than having one neuron in the output layer, have N binary neuron leading to multiclass classification. Classes are converted into binary indicators for N binary neurons. Tuning parameters are size and decay. The model is sensitive to highly correlated predictors and near-zero predictors. The data must also be centered and scaled.

For modeling with a neural network, there are several variations that can impact the effectiveness of the model. These include the learning rate, the number of layers in the neural network, and the number of nodes in each layer. Several models were investigated changing these parameters. Additionally, we attempted to use different variables in the modeling.

- 5 nodes, 1 layer, 0.01 learning rate, no soil predictors, random sample -> Kappa 0.56, Accuracy 0.73
- 5 nodes, 1 layer, 0.01 learning rate, soil predictors, random sample -> Kappa 0.57, Accurarcy 0.74
- 9 nodes, 1 layer, 0.01 learning rate, soil predictors, random sample -> Kappa 0.59, Accuracy 0.75
- 5 nodes, 1 layer, 0.01 learning rate, soil predictors, tiered sample -> Kappa 0.60, Accuracy 0.70

The winning neural network model is the one with the highest Kappa with the tiered sampling.

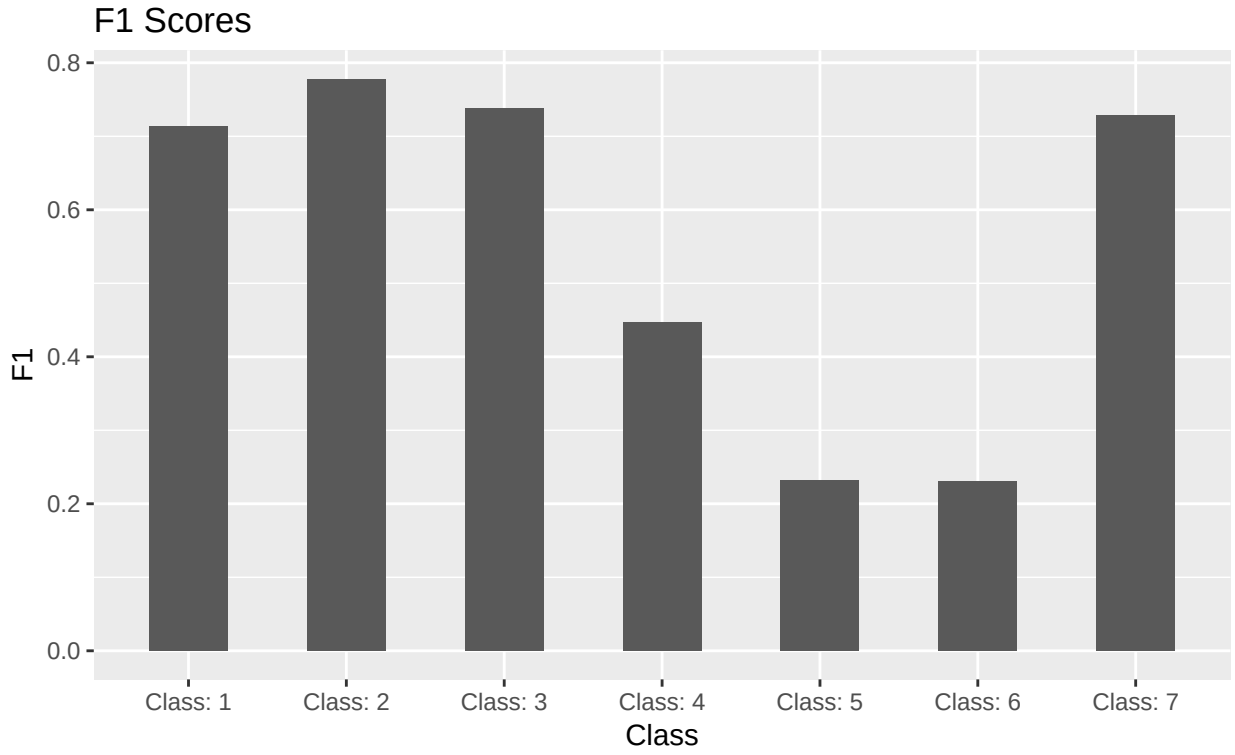


Figure 14: F1 for Neural Network

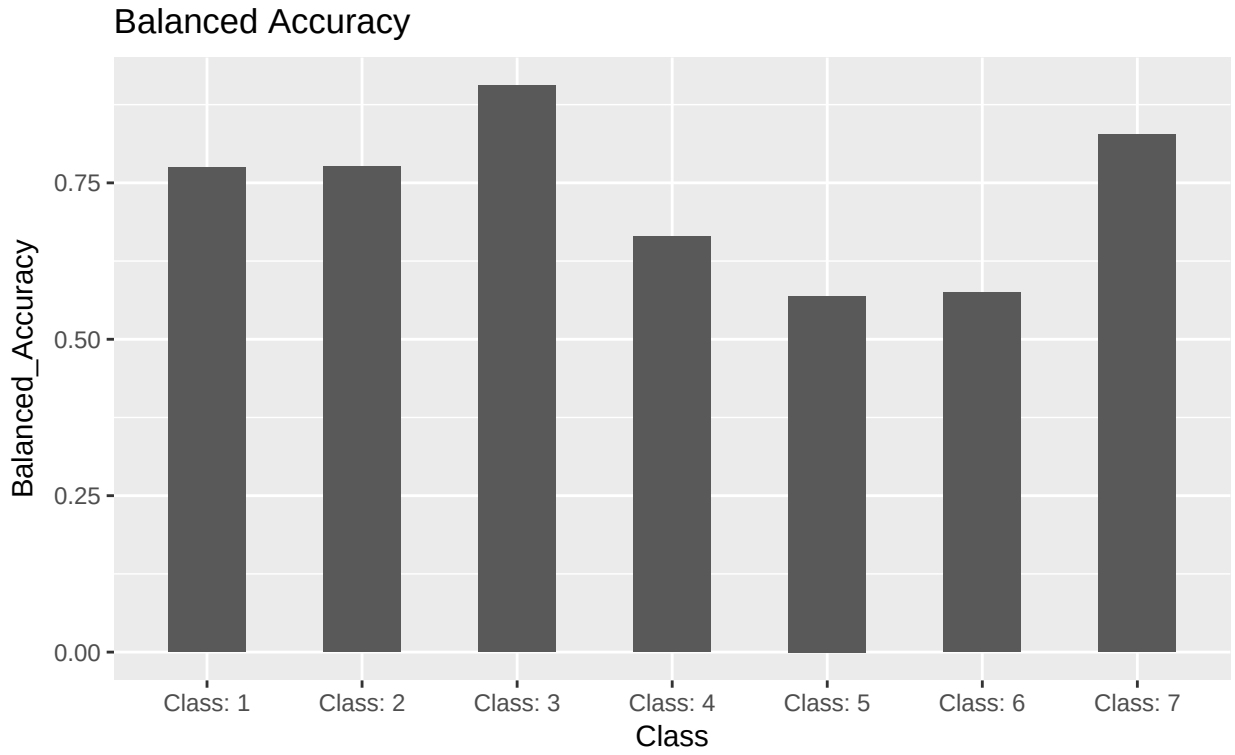


Figure 15: Balanced Accuracy for Neural Network

7.3 Random Forests

Each tree in the forest casts a vote for the classification of a new sample, and the proportion of votes in each class across the ensemble is the predicted probability vector. Tuning parameter is mtry, the number of variables randomly sampled as candidates at each split. There is no need to preprocess the data.

7.3.1 Random Forest - Reduced Feature Set, Stratified Sample, 5-fold cross-validation and 3 repeats, random search for mtry

7.3.1.1 Sample

For our first Random Forest model, we used a stratified sample of 23,243 observations for training. This amounted to 5% of the observation left after removing 20% of total observation for use in backwards recursive feature selection. Stratified sampling maintains the class distribution found in the original data set.

7.3.1.2 Predictors

Our original data set was augmented with variables suggested by our exploratory data analysis, including linear distance variable was also created using the vertical and horizontal distance to hydrology variables and interactions between our Areas and Elevation, as suggested by the distinct distributions of the areas by elevation. We used recursive backwards feature selection on 20% of our full data set with 'Kappa' set as our metric to maximize. The algorithm is configured to explore all possible subsets of the attributes. It returned 16 predictors.

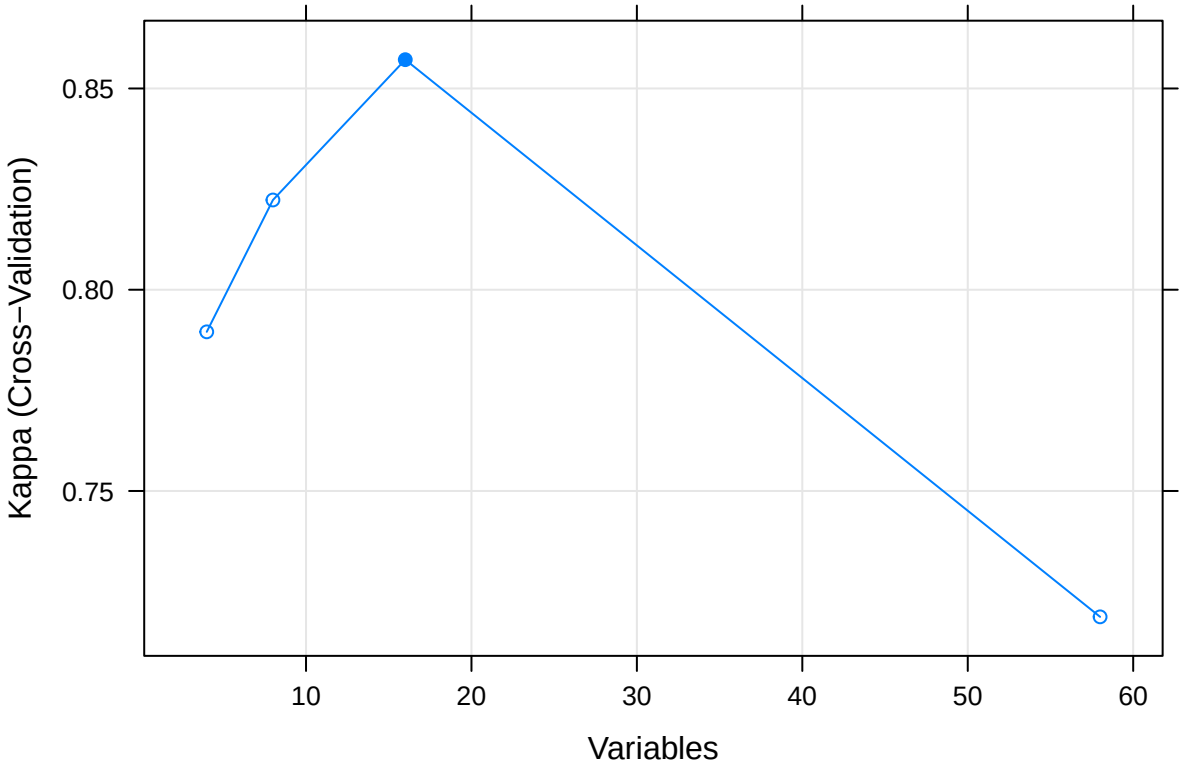
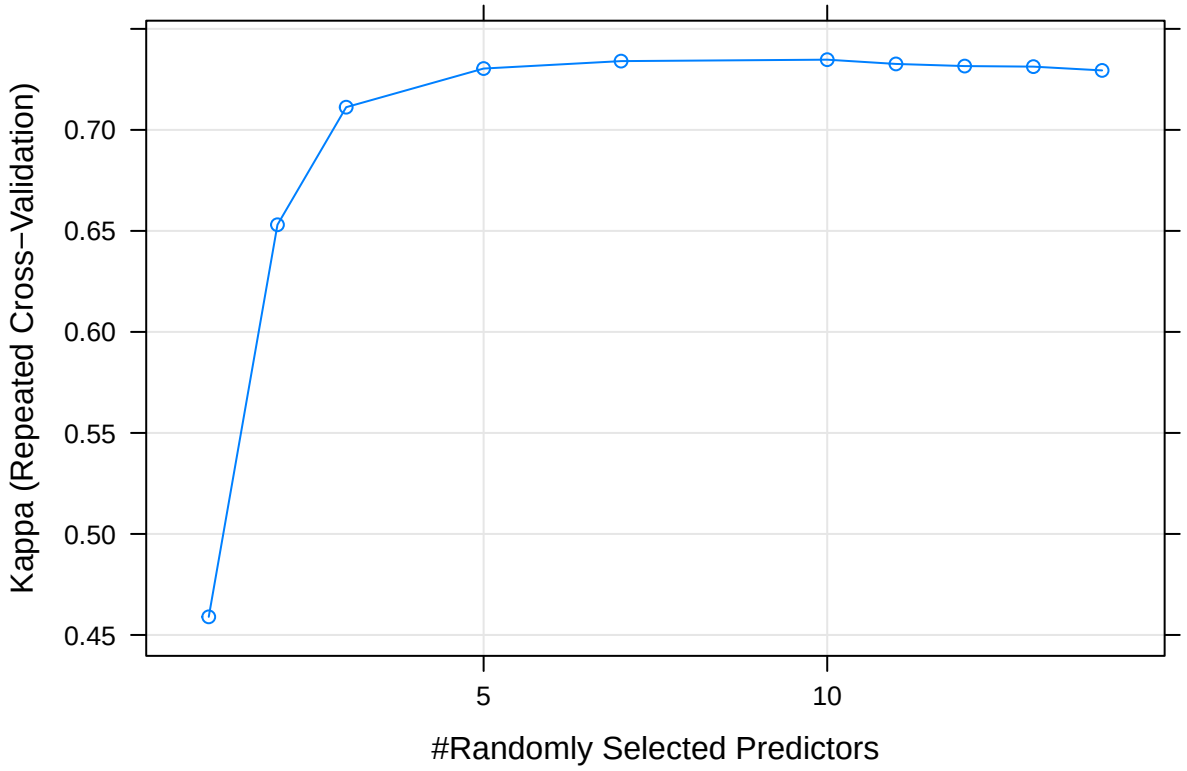


Table 7: RFE Predictors

rfe.predictors
Elevation
HDist.FirePoint
HDist.Roadway
trans.LDist.Hydrology
VDist.Hydrology
Hillshade.3pm
Hillshade.12pm
Aspect
HDist.Hydrology
interElAr3
Slope
interElAr1
SoilType22
SoilType23
SoilType2
SoilType33

7.3.1.3 Tuning Methodology

In order to get better accuracy from our algorithm we used algorithm tuning to find the best parameter value for our problem. We used random search strategy to try random values within a range. Kappa was used to select the optimal model using the largest value. The final value used for the model was mtry = 10.



7.3.1.4 Results

1	2	3	4	5	6	7
136109	19850	2	0	293	34	3759
23792	192703	2012	4	4138	1889	167
7	1412	23552	818	161	3177	0
0	40	239	1206	0	108	0
48	340	9	0	2602	2	0
38	828	1358	59	20	7988	0
1004	135	0	0	0	0	11661

The model had an accuracy of 85.1%, but accuracy is not a good measurement for our imbalanced data set since it reflects the underlying class distributions. We will use alternative metrics such as Kappa, Balanced Accuracy, and F1 score to evaluate our models. Our model achived a Kappa value of 75.7 and had poor results in our underrepresented classes.

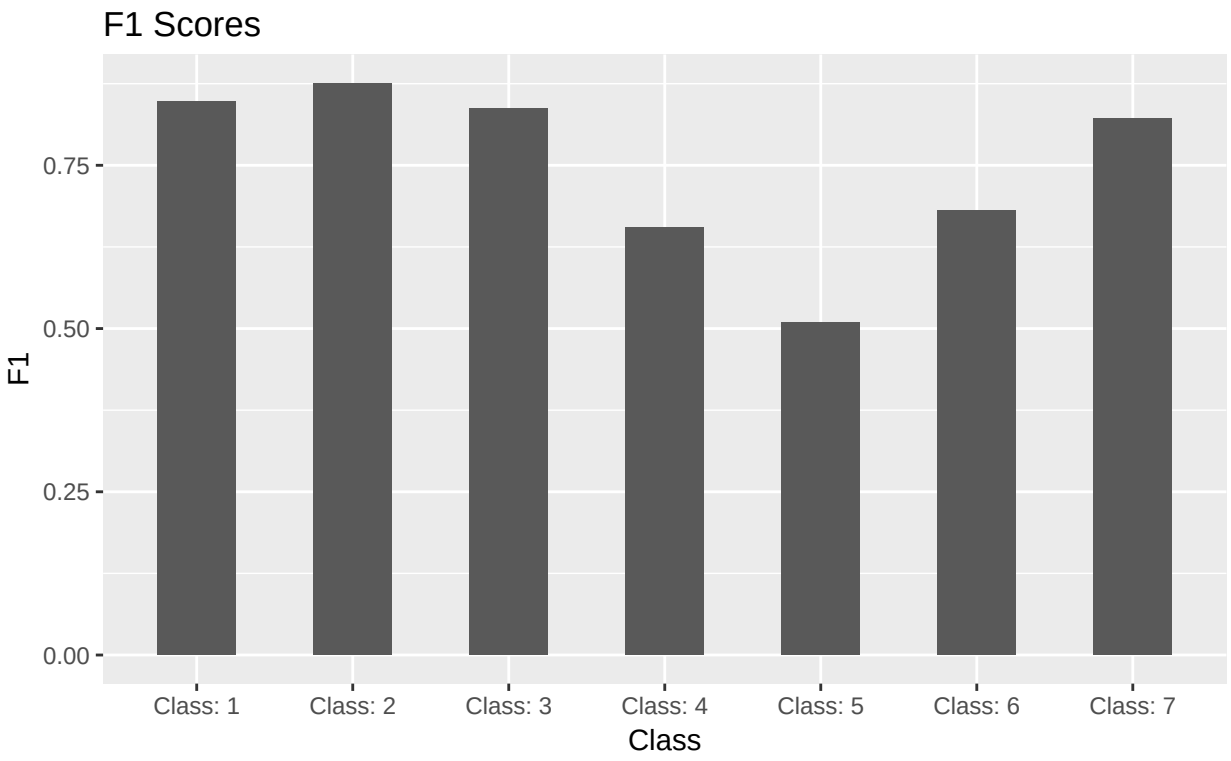


Figure 16: F1 for Random Forest Model #1

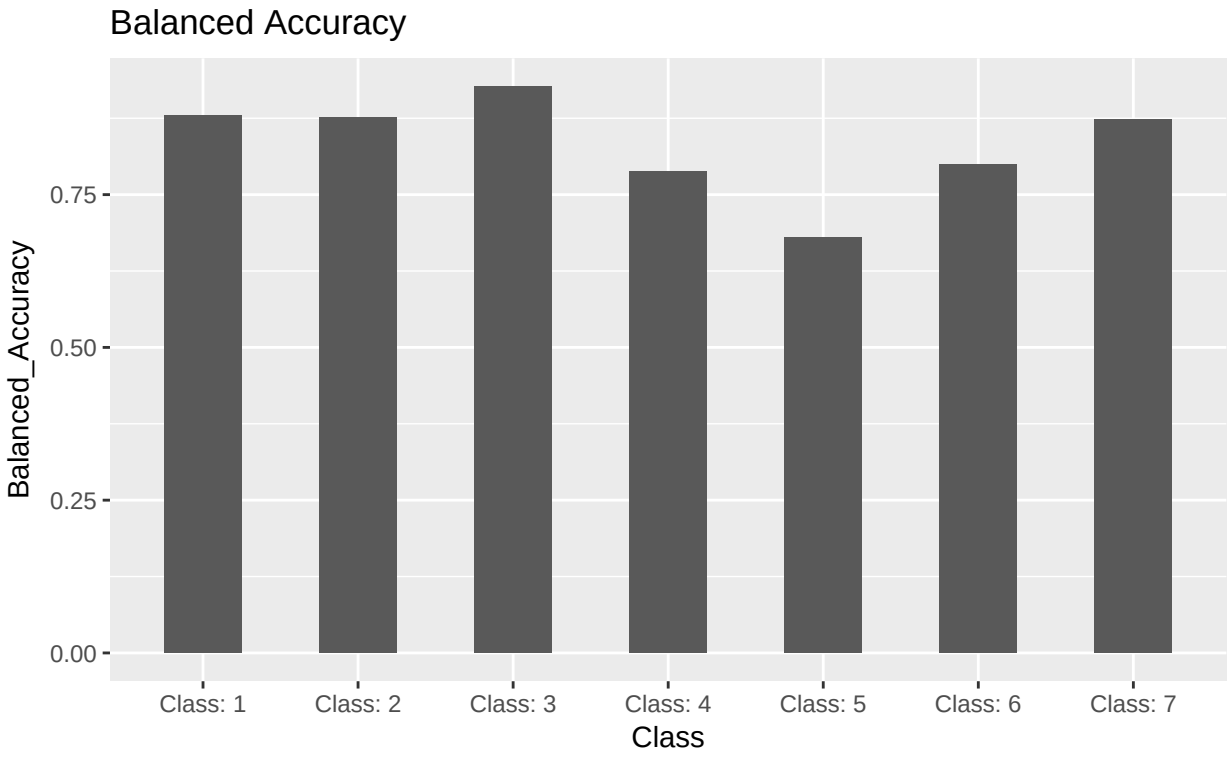


Figure 17: Balanced Accuracy for Random Forest Model #1

7.3.2 Random Forest - PCA Soil, Stratified Sample, 5-fold cross-validation and 3 repeats, grid search for mtry

7.3.2.1 Sample

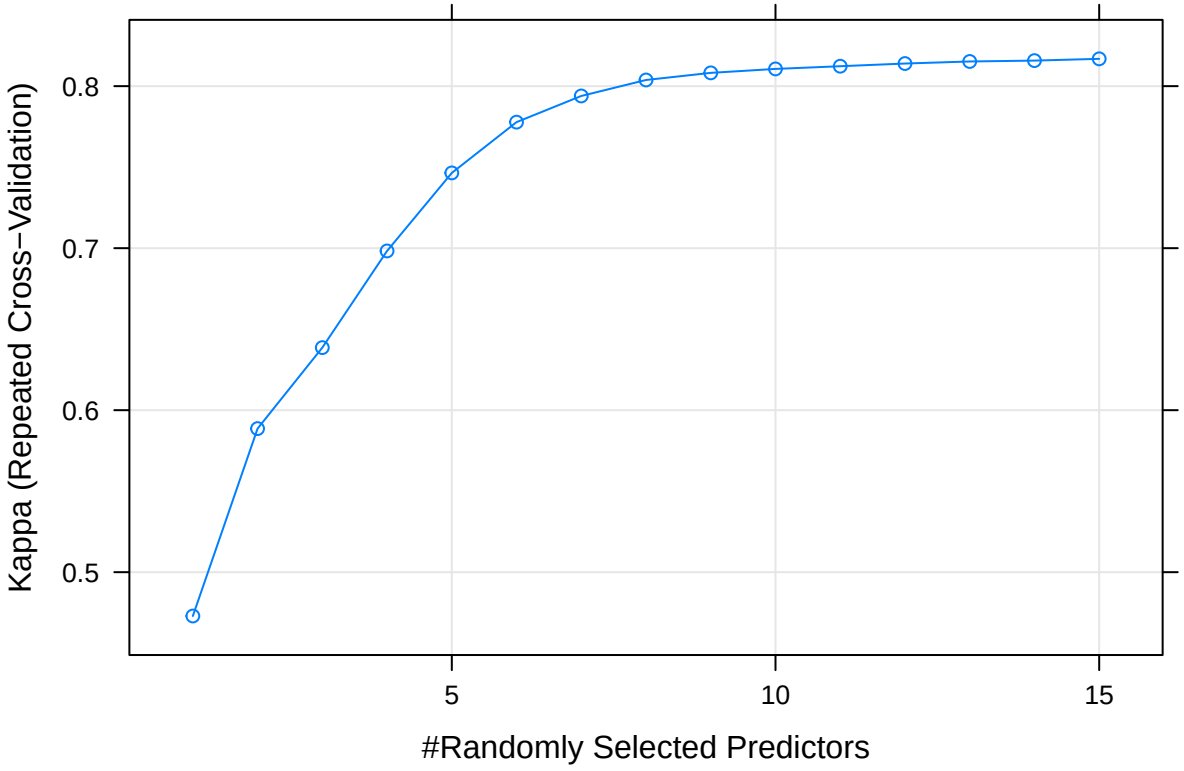
For our second Random Forest model, we used the stratified sample of 46,484 observations for training.

7.3.2.2 Predictors

For our second Random Forest model we removed the 40 soil values and replaced them with 16 variables that explained 90% of the variance.

7.3.2.3 Tuning Methodology

We used grid search strategy to try random values within a range. Each axis of the grid is an algorithm parameter, and points in the grid are specific combinations of parameters. Because we are only tuning one parameter, the grid search is a linear search through a vector of candidate values. Kappa was used to select the optimal model using the largest value. The final value used for the model was mtry = 15.



7.3.2.4 Results

1	2	3	4	5	6	7
135746	10408	6	0	161	48	1659
15973	191927	902	0	2378	858	140
16	741	23807	396	113	1821	0
0	7	177	1542	0	87	0
53	332	31	0	4160	9	3
30	504	819	39	22	9680	0

1	2	3	4	5	6	7
706	57	0	0	0	0	12965

The model had an accuracy of 90.8%, achieved a Kappa value of 85.1% and had much improved results in our underrepresented classes.

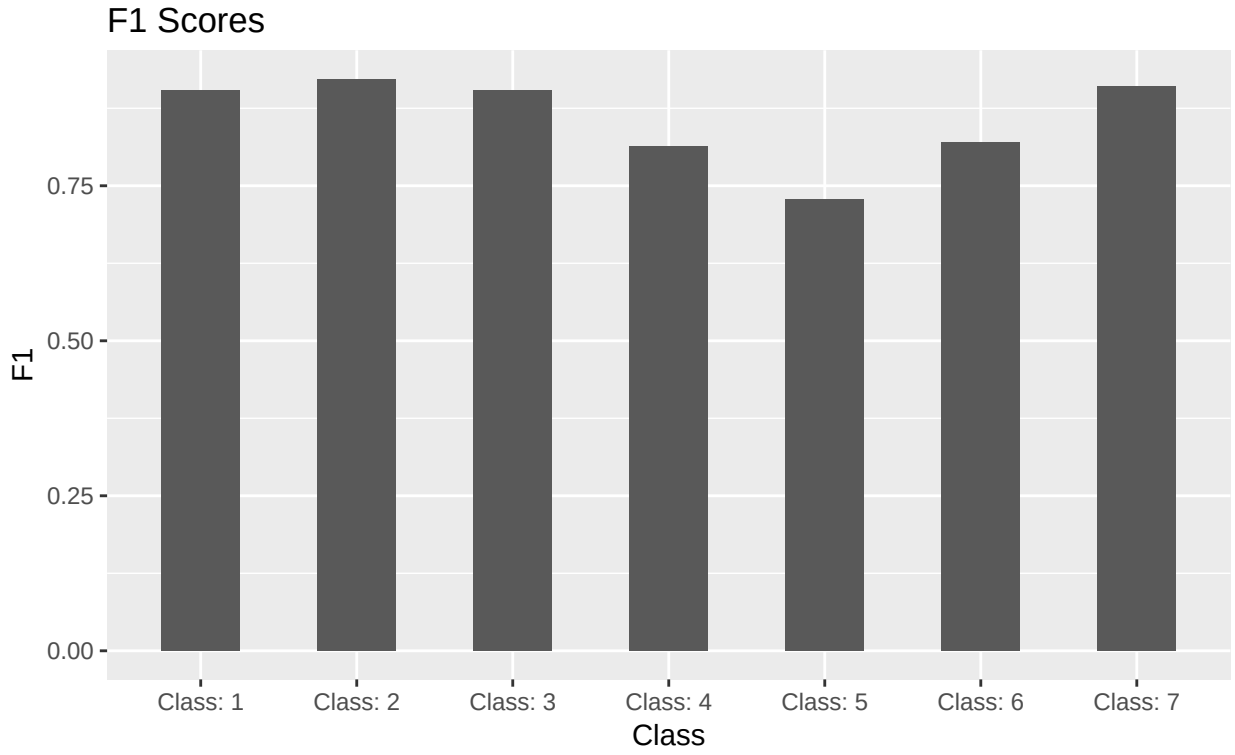


Figure 18: F1 for Random Forest Model #2

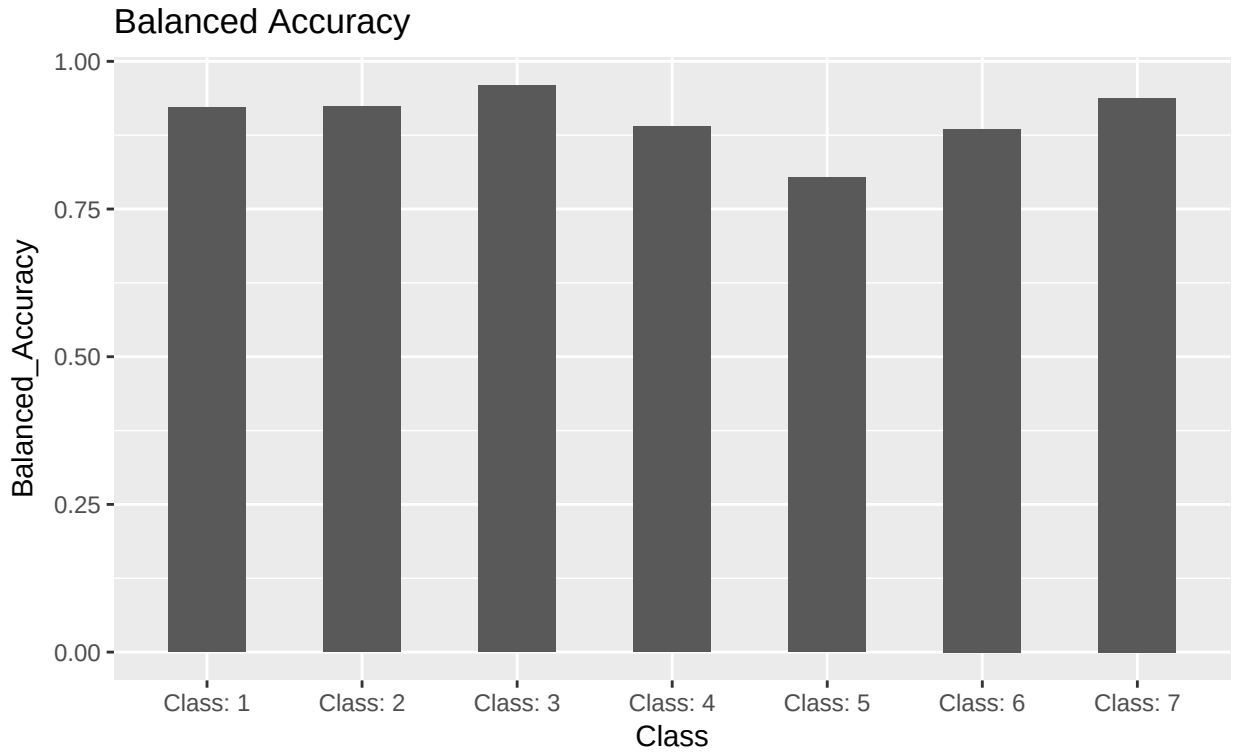
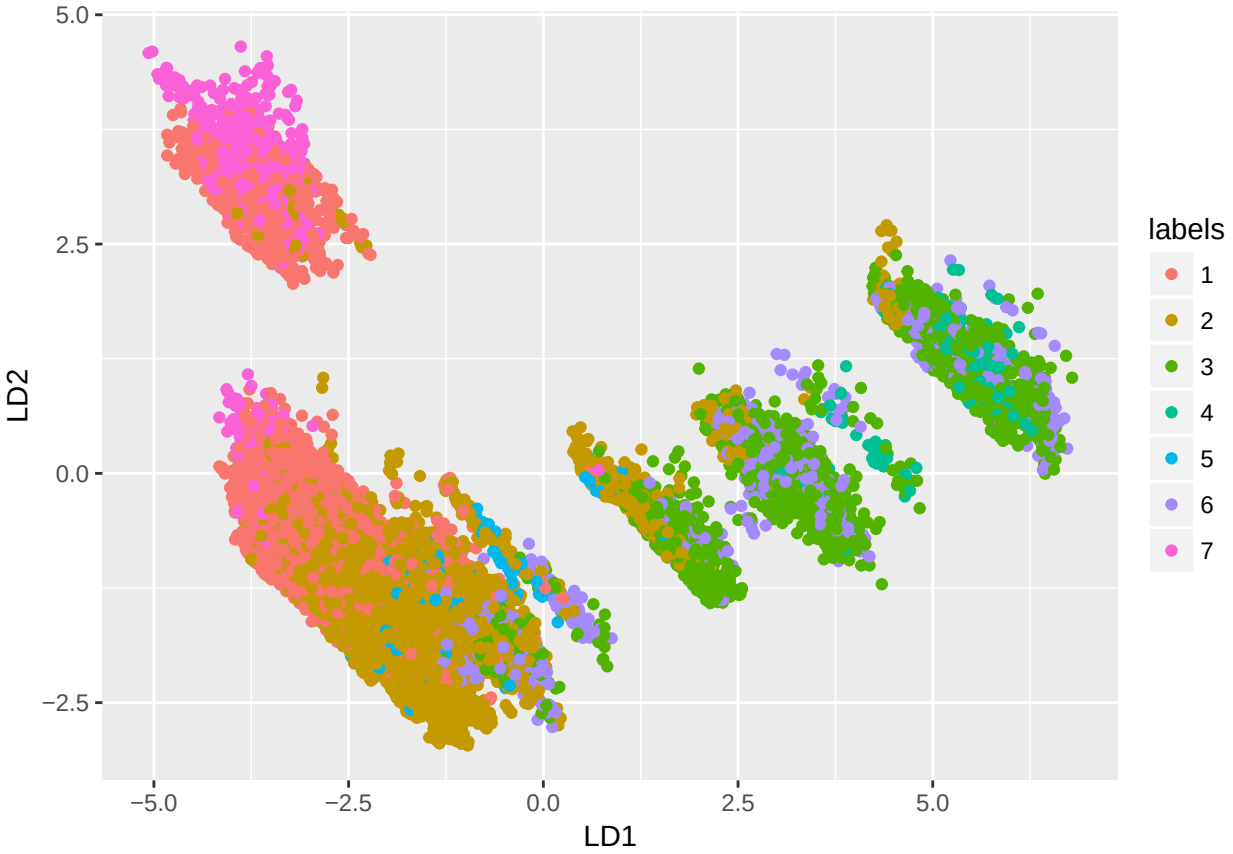


Figure 19: Balanced Accuracy for Random Forest Model #2

7.4 Discriminant Analysis

LDA assumes normal predictors and that the predictors have equal variance. LDA then estiamtes the mean and variance for the predictors for each class. There is no tuning parameter and the model is sensitive to highly correlated predictors and near-zero predictors. The data must also be centered and scaled.



7.5 K-Nearest Neighbors

Predictions are made for by searching through the entire training set for the K most similar instances and summarizing the most common class for those K instances. KNN is suited for lower dimensional data. We will need to center and scale predictors prior to performing KNN. The model is sensitive to near-zero predictors and the tuning parameter is k.

K Nearest Neighbor (KNN) modeling involves a relatively simple algorithm that classifiers a target variable based on the classification of the "k" nearest neighbors. K is selected using cross-validation to avoid over-fitting. For our model, we used five- fold cross validation across K values ranging from 2-10. Predictor variables were normalized across a random sample of 40,000 training observations. The best performing model used three Shade PCA variables along with summarized Soil Types in the form of Forest Zones. An optimal "k" value of 3 was elected for this model, which resulted in model that performed poorly with a 44.1% misclassification rate.

7.6 Logistic Regression

The model is sensitive to highly correlated predictors and near-zero predictors. The data must also be centered and scaled.

7.7 Support Vector Machine

A support vector machine algorithm an optimal hyperplane which categorizes new examples into a class. The tuning parameter is C, allows violation of the margins with C=0 allowing for no violation and higher variance

and larger C resulting in a less sensitive algorithm. the The model requires centering and scaling of data.

8 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. Principal components analysis helped us narrow our soil variables down from 40 to 16 variables. Additionally, we prepared the data by creating a subset of 70% of the data to train our models on, and a validation set of 30% of the data to test those models on.

Next we experimented with different samples, feature sets and several classification algorithms and assessed our initial modeling results. Our next step will be to conclude our modeling methods and compare our results.