Team 4 – Bryn Wiley, Armandas Bartas, and Dmitri Arykov

Section 1: Abstract

The goal of this project was to design a predictor for the forest cover type in Roosevelt National Forest of Northern Colorado, in 30x30m observational plots, from various observed cartographic variables. Additionally, we aimed to gain a relative understanding of the ranking of values for important variables with respect to each cover type (their "niche spaces"). Of the explanatory variables, the variables most important for prediction were elevation. soil type, and various distance measurements to significant features from sample sites. The methods applied were logistic regression, tree based gradient boosting, classification trees, and random forest. Random forest provided the lowest overall CV misclassification rate at 12.3%, but gradient boosting may be better for prediction in some scenarios. Naturally, of the seven cover types, some were easy to differentiate, while others proved more difficult. Specifically, Spruce/Fir and Lodgepole pine, and Douglas Fir and Ponderosa Pine, were often mistaken for each other. Overall, this report describes effective predictive methods for forest cover types in this region, and provides insight into the habitat preference of the trees in these cover types.

Section 2: Variables & Transformations.

In this section, we will examine the response and explanatory variables, and provide summary statistics.

The response variable for this dataset was cover type, or the predominant type of tree present in a 30x30m square observation plot. There were seven types of cover type, registered in the dataset as integers representing a categorical variable for classification.

1 - Spruce/Fir, 2 - Lodgepole Pine, 3 - Ponderosa Pine, 4 - Cottonwood/Willow, 5 – Aspen, 6 - Douglas-fir 7 - Krummholz

The dataset provided (15120 observations) contains both various cartographical features and the cover type (Cover_Type). There was an equal number of observations for each class. (2160 for each class)

Variables:

Elevation - Elevation in meters
Aspect - Aspect in degrees azimuth (degrees on a compass)
Slope - Slope in degrees
Horizontal_Distance_To_Hydrology - Horizontal Distance to nearest surface water features in m
Vertical_Distance_To_Hydrology - Vertical Distance to nearest surface water features in m
Horizontal_Distance_To_Roadways - Horizontal Distance to nearest roadway in m
Hillshade_9am (0 to 255 index) – Hill shade index at 9am, summer solstice (amount of light, no units)
Hillshade_Noon (0 to 255 index) – Hill shade index at noon, summer solstice (amount of light, no units)
Hillshade_3pm (0 to 255 index) – Hill shade index at 3pm, summer solstice (amount of light, no units)
Horizontal_Distance_To_Fire_Points - Horizontal Distance to nearest wildfire ignition points, in m
Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation
Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation
Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation, the response variable

The wilderness areas are: 1 - Rawah Wilderness Area, 2 - Neota Wilderness Area, 3 - Comanche Peak Wilderness Area, 4 - Cache la Poudre Wilderness Area. These are large sections of the park, which divide it into four sections. There were 40 different types of soil measured, and the list may be viewed in the appendix. These types were classified based on content and texture.

Wilderness area and the soil type were presented as binary (sparse) data. For clarity, the soil type variables were combined into one soil type variable with 38 factor levels (as some soil types are missing in the dataset), The soil type was then transformed by selecting only the top 20 soil types (by frequency), with the rest assigned the same "unclassified" type. This was done to improve discrimination, as some soil types were very rare. This "dropped" soil type was used for all models in place of soil type, except in the random forest model. The wilderness area variables were also presented as sparce data, and were also combined into one with 4 different factor levels, according to the wilderness area type.

Aspect was recorded in degrees of azimuth, or as degrees on a compass (ie: 0 is north, 180 is south). We noted that minimum of the aspect variable was at 270 degrees for many cover type classes. As such, one transformation applied was to rotate the aspect 270 degrees, so that 0 coincides with the minimum, potentially providing a more normal looking distribution. Another transformation of the aspect variable applied was binning into cardinal directions, with binning as follows: factor N for values between 338.75 and 22.5 degrees, NE for values between 22.5 and 67.5 degrees, E between 67.5 and 112.5 degrees, SE for values between 112.5 and 157.5 degrees, S for values between 157.5 and 202.5 degrees, SW for values between 202.5 and 247.5 degrees, W for values between 247.5 and 292.5 degrees, NW for values between 292.5 and 337.5 degrees. The results of the transformation can be viewed in Figure 1.

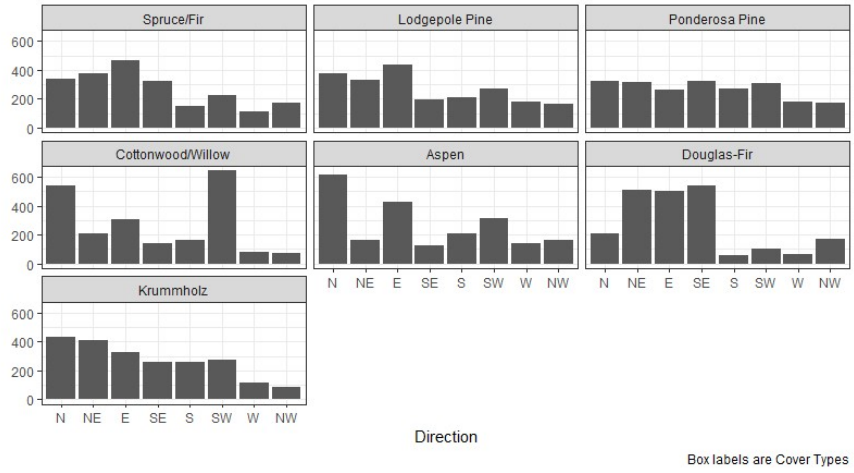

Figure 1. The binned aspect variable, with bin frequency plotted for each cover type.

Horizontal distance measurements showed more skewed distributions, so we applied log transforms to those measurements in an attempt to make the distributions look more normal. Vertical distance to hydrology was cube rooted instead because it contained negative values. Additionally we binned the distance measurements (both horizontal and vertical) into categories with 5 different levels based on dataset quantiles, as an alternative method of reducing variable skew.

Some variables exhibited visible correlation. This is demonstrated most clearly with correlation between hillshade measurements, as shown in Figure 2, and between hillshade measurements and aspect, as shown in Figure 3.

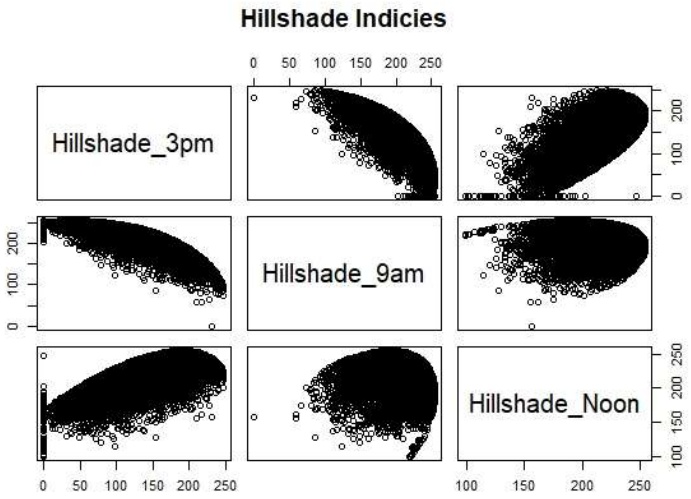

Figure 2: Bivariate summaries between the three hillshade measurements

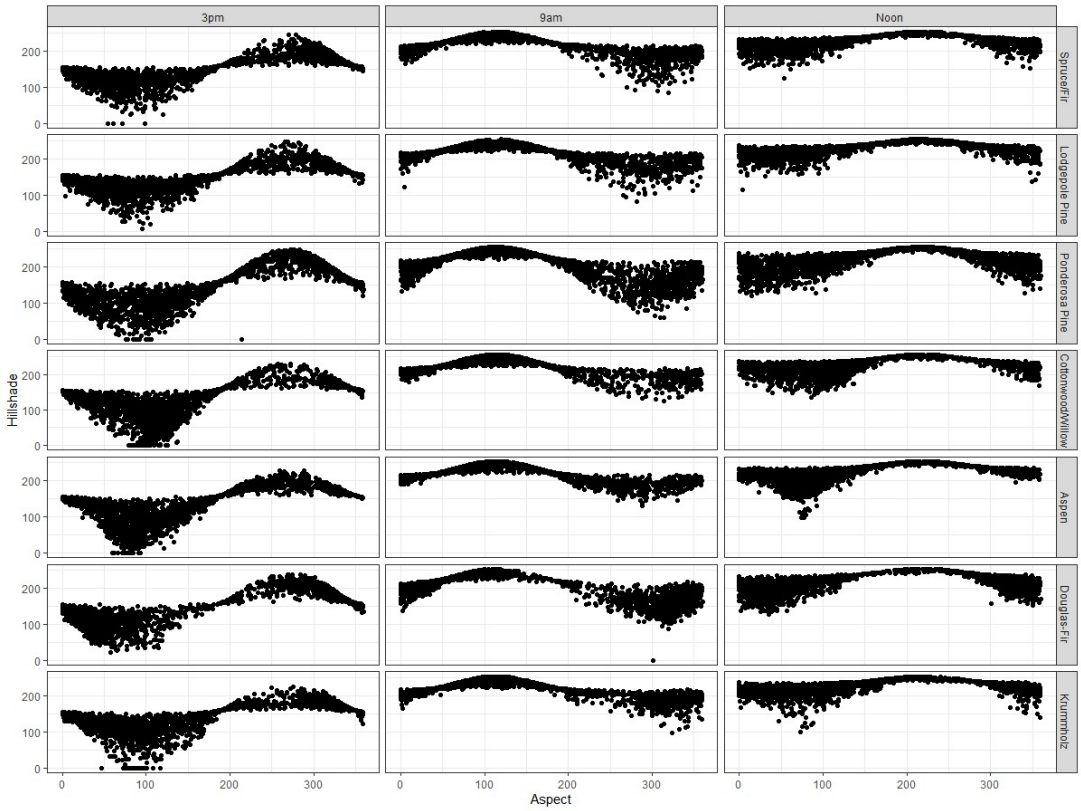

Figure 3: Bivariate summaries between the three hillshade variables (represented on the y axis, with the three separate variables represented in x-axis box labels), and the Aspect variable (represented on the x-axis), for each class (represented with the y-axis box labels).

To attempt to eliminate the correlation between hillshade observations, we applied an averaging transformation, by averaging the three hillshade variables to create a new hillshade average variable.

From visual analysis, the transformed soil type and elevation appeared to be the best visible predictors, as demonstrated in figures 4 and 5 with clear class separation. The other variables displayed much more subtle differences between classes.

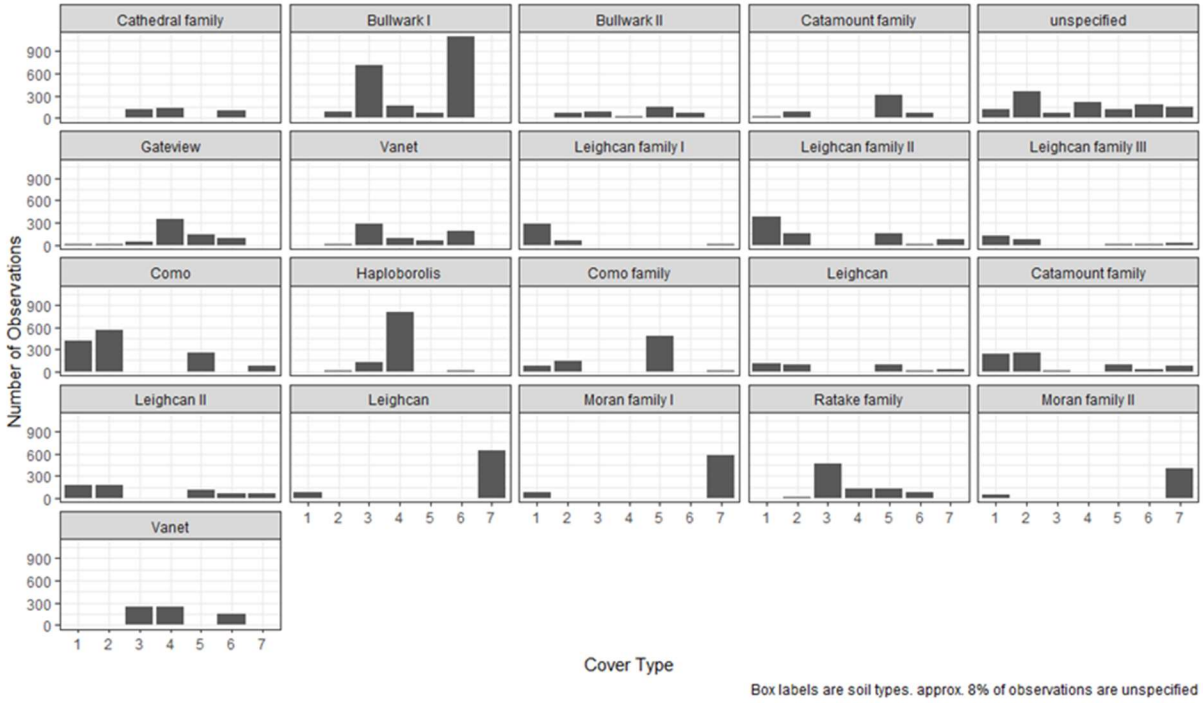

Figure 4: The frequency (represented on the y axis) of each soil type (represented in box label) for each cover type (represented on the x axis as integer labels).

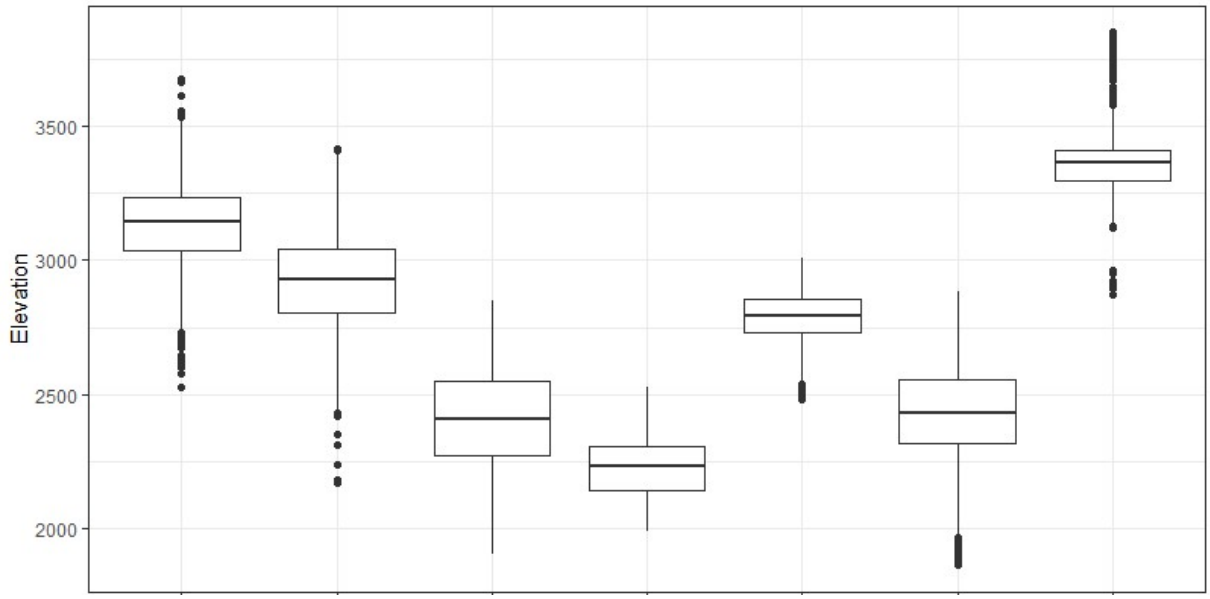

Figure 5: The distribution of the elevation variable plotted for each class (on the x-axis). Elevation is measured in m. This graph displays clear class separation, the clearest of all features.

Section 3: Model analysis, Comparison, and Interpretation

In this section, we will examine the top models produced by various classification approaches, and their performance results. We will then compare the models to find the best predictor and provide interpretations.

Logistic Regression

To attempt to find a subset of important variables for prediction, and also to facilitate interpretation, we applied multinomial logistic regression to this dataset using the vglm function in R. Typically, feature selection can be accomplished through stepwise variable selection. However, using this dataset and regression software, stepwise variable selection was prohibitively slow for large numbers of features, and often failed to eliminate any variables from the model (using both AIC and BIC as selection criteria). As such, we used a two part strategy. First, we fit combinations of logically grouped features to a training set, and measured performance on a holdout set. For brevity, the results of this step are not included in this report. Then, we took the best performing combinations of variable "sets", performed stepwise variable selection with BIC, and measured prediction performance using five fold CV.

Model (described in caption)	Spruce/ Fir	Lodgepole Pine	Ponderosa Pine	Cottonwood/ Willow	Aspen	Douglas -Fir	Krummholz	Overall
A	0.340	0.463	0.455	0.131	0.230	0.352	0.113	0.298
B	0.374	0.493	0.580	0.196	0.285	0.375	0.121	0.347
C	0.343	0.447	0.426	0.010	0.240	0.330	0.105	0.284
D	0.352	0.475	0.398	0.107	0.217	0.334	0.103	0.283
E	0.333	0.471	0.403	0.093	0.211	0.316	0.093	0.274

Table 1: Misclassification rates for each class and overall, from five-fold CV. Model A uses all untransformed variables included in the dataset. Model B uses only Elevation, Soil Type, and Wilderness Area, denoted as the "base". Model C is the base, along with log/cube transformations of distance measurements, the hillshade measurements, and the binned transformation of the aspect variable. Model D is the base, along with binned transformation of the distance variables and their interaction with elevation. Model E is the "best", and includes the same variables as Model D, along with slope, binned aspect, and the interaction between slope and elevation.

Model (described in caption)	Spruce/ Fir	Lodgepole Pine	Ponderosa Pine	Cottonwood/ Willow	Aspen	Douglas -Fir	Krummholz
A	1.990	2.119	2.098	1.683	2.023	2.192	1.420
B	2.027	2.138	2.261	1.969	2.125	2.350	1.454
C	1.986	2.114	2.044	1.489	2.004	2.131	1.405
D	1.997	2.107	2.064	1.502	2.019	2.121	1.377
E	1.974	2.097	2.007	1.391	1.983	2.063	1.378

Table 2: 80% prediction interval widths for each class, from five-fold CV. Model A uses all untransformed variables included in the dataset. Model B uses only Elevation, Soil Type, and Wilderness Area, denoted as the "base". Model C is the base, along with log/cube transformations of distance measurements, the hillshade measurements, and the binned transformation of the aspect variable. Model D is the base, along with binned transformation of the distance variables and their interaction with elevation. Model E is the "best", and includes the same variables as Model D, along with slope, binned aspect, and the interaction between slope and elevation.

The CV evaluation results for different models are shown in table 1 and table 2, with misclassification rates and 80% interval widths respectively. Although not the best performing model, the model including all original untransformed features (except for Soil Type) is included for comparison (Model A in table). Similarly, visual analysis identified Elevation, Soil Type, and Wilderness Area as strong predictors. As well, attempts at stepwise variable selection frequently identified these variables as strong predictors, significantly increasing AIC upon their removal. As such, we identified this subset as a set of strong predictors, denoted as the "base", and also performed CV analysis on them (Model B in table). To attempt to find the overall best predictor, we also combined all of the variables in the top performing models analysed (some note included here), performed stepwise variable selection, and performed CV analysis to provide the "best" model. Although not part of this process, we also fit a model with log transformed distance measurements, hillshade measurements, the binned transformation of the aspect variable, and the "base" variables. Unexpectedly, this model had a relatively high CV performance (model C in table).

The three best performing models (of those analysed) shown in table 1 and 2 are as follows. The third place was the unexpected model fitted for interpretability, as described above. The second place model included the "base" variables (Elevation, Soil Type, and Wilderness Area), alongside the binned transformation of the four distance measurements and the interaction terms between elevation and distance to roadways and between elevation and distance to fire points (both interaction distances binned). This is model D in the tables. Finally, the top performing model was the "best" model, as described above. After variable selection, this "best" model included the same variables as the second best performing model, but also the binned hillshade measurements, slope, and interaction between elevation and slope. This is model E in the tables.

Overall, these results validate our conclusions that elevation, soil type, and distance measurements are important predictors. CV analysis on multinomial logistic models that did not include distance measurements were worse performing. As well, inclusion of hillshade and slope variables in the "best" predictor improved CV misclassification error by less than a percent over the second best model, again validating the importance of the distance measurements as important predictors.

Similar to other models examined in this report, and as shown in table 1, the categories Spruce/Fir, Lodgepole Pine, Ponderosa Pine, and Douglas-Fir showed a disproportionately high misclassification rate. An analysis of the 50% confidence intervals (see appendix table A1) shows that Spruce/Fir is most often misclassified as Lodgepole Pine, with the same relationship between Lodgepole Pine and Spruce/Fir, between Ponderosa Pine and Douglas-Fir, and between Douglas Fir and Ponderosa Pine.

By analyzing the coefficients of the third best logistic model, we can generate interpretations about the relative distance measurements and elevation, two important variables for prediction, for each class. Table 3 shows the relative order of expected distance measurements and elevation by the coefficients (largest to smallest coefficients), all other variables being equal. These will be further discussed in the conclusion.

Variable	**Relative Value per Cover Type (Decreasing)**
Horizontal Distance to Hydrology	Ponderosa Pine, Lodgepole Pine, Douglas-Fir, Aspen, Spruce/Fir, Krummholz, Cottonwood/Willow
Vertical Distance to Hydrology	Cottonwood/Willow, Ponderosa Pine, Douglas-Fir, Aspen, Lodgepole Pine, Spruce/Fir, Krummholz

Horizontal Distance to Fire Points	Krummholz, Cottonwood/Willow, Spruce/Fir, Douglas-Fir, Lodgepole Pine, Aspen, Ponderosa Pine
Horizontal Distance to Roadways	Cottonwood/Willow, Douglas-Fir, Lodgepole Pine, Spruce/Fir, Ponderosa Pine, Krummholz, Aspen
Elevation	Krummholz, Spruce/Fir, Lodgepole Pine, Aspen, Douglas-Fir, Ponderosa Pine, Cottonwood/Willow

Table 3: Relative expected value of variables for each cover type, implied from multinomial logistic regression coefficients.

Gradient Boosting

In order to gain further insight into variable importance and improve prediction performance, but not provide additional interpretation, we applied tree-based gradient boosting using the xgboost library in R. We used a two-step process to find the best model for this task. First, we compared different models with all available features, but with different combinations of transformations, using training/holdout splits and measuring out of sample performance. This was done using the default parameter values for xgboost. Second, we used the best variable transformations, along with 5-fold cross validation, to choose the best tuning parameters for the xgboost function. We were limited by computing power, so we were only able to choose from a small range of values for each parameter.

For clarity, the results of the first step are not included in this report. However, they suggested that, unlike other models examined, any transformation of the distance measurements hindered classification based on out of sample misclassification rate. Similarly, unlike in other models, taking the average of the three hillshade measurements (at 9am, noon, and 3pm) to replace them with a single variable for each observation improved out of sample misclassification rate. The other transformation that improved out of sample prediction performance was binning the aspect variable.

With these transformations, we used CV to select the eta, gamma, nround, and max_depth parameters of the xgboost function. The range of values for each was 0.1, 0.3, or 0.5 for eta, 0, 3, or 6 for gamma, 100, 300, or 500 for nround, and 6, 9, or 12 for the max_depth parameter. The lowest misclassification rate was found with values of 0.3, 0, 500, and 12 for eta, gamma, nround, and max_depth respectively. As well, this set of parameters gave the second lowest variance in CV misclassification error across the separate classes.

	Spruce/ Fir	**Lodgepole Pine**	**Ponderosa Pine**	**Cottonwood/ Willow**	**Aspen**	**Douglas-Fir**	**Krummholz**	**Overall Misclassification Rate**
Misclassification Rate	0.230	0.284	0.165	0.031	0.045	0.107	0.033	**0.128**
80% Interval Length	1.312	1.383	1.258	1.065	1.153	1.218	1.073	
50% Interval Length	1.011	1.019	1.12	1.00	1.012	1.013	1.00	

Table 4: Misclassification rates and 80% and 50% interval lengths from 5-fold CV for the best-performing gradient boosting model. The overall CV misclassification rate is also given.

Using these parameters, we achieved the CV misclassification rates, 80% and 50% interval lengths, and overall CV misclassification rate shown in table 4. Uneven classification rates were present, with Spruce/Fir, Lodgepole Pine, and Ponderosa Pine displaying a disproportionately high misclassification

rate. These classes also displayed the same pattern of misclassification shown by the logistic model, namely Spruce/Fir misclassified most often as Lodgepole Pine and vice versa, and Ponderosa Pine misclassified most often as Douglas-Fir and vice-versa.

However, using this model we also confirmed the high importance of elevation, soil type, and the distance measurements for prediction. Figure 6 shows the variable importance graph generated using the xgboost library. From this, we see that elevation, the distance measurements, and one of the soil types are the top 5 most important variables for prediction.

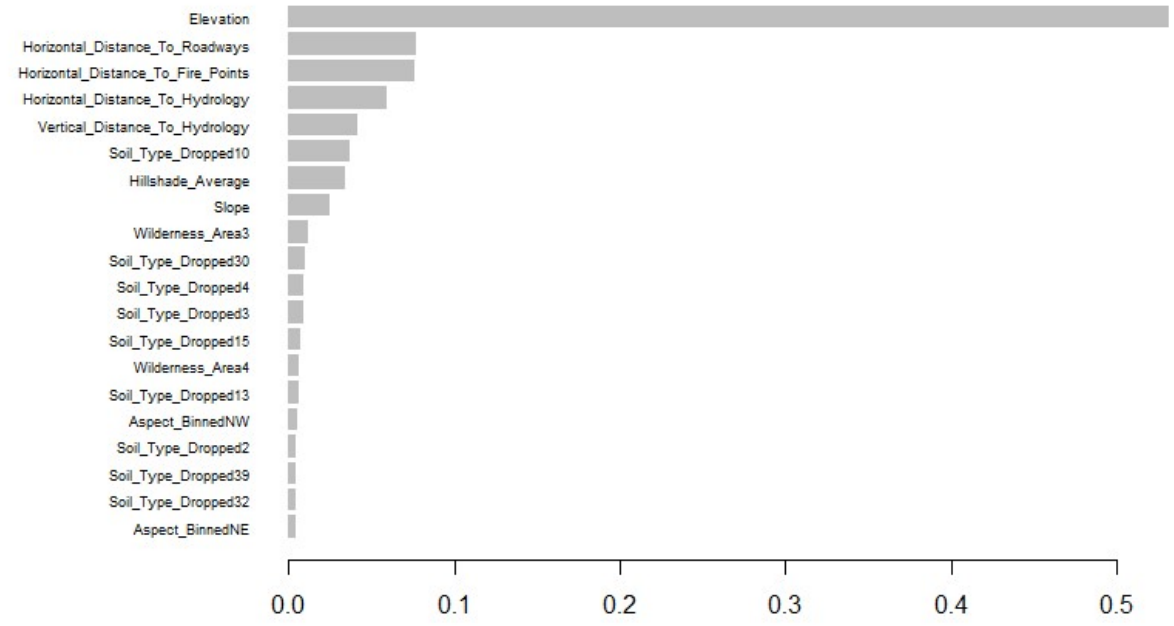

Figure 6: The relative importance of the top twenty variables by the xgboost measure of variable importance, for the best performing gradient boosting model. The remaining twenty variables are eliminated from this graph for readability.

Classification Trees

To provide the clearest interpretation, and also a possible good predictor, we applied classification tree models to this dataset with RPart. To determine the best model, we ran an experiment to evaluate the average misclassification rate with 5 fold CV for every possible combination of variables on two datasets: the original dataset with all possible log/cube root transformations, and the original dataset with all possible binning transformations. Both datasets also used the transformed soil type feature. The best performing formulas at each level of variable inclusion were then extracted. Table 5 shows an example output with the top 3 models using 6 variables, using all possible log/cube root transformations.

Misclassification rate	**Explanatory variables**
0.347	Elevation , Aspect, Wilderness_Area, log_Horizontal_Distance_To_Fire_Points, log_Horizontal_Distance_To_Roadways, cube_Vertical_Distance_To_Hydrology

0.351	Elevation, Aspect, Hillshade_3pm, Wilderness_Area, log_Horizontal_Distance_To_Fire_Points, cube_Vertical_Distance_To_Hydrology
0.353	Elevation, Aspect, Hillshade_3pm, Wilderness_Area, log_Horizontal_Distance_To_Roadways, cube_Vertical_Distance_To_Hydrology

Table 5: The top three models and their 5-fold CV misclassification rate, including only 6 features and using log or cube root transformations where applicable.

Figures 7 and 8 show the CV misclassification rate of best performing formula for each number of variables included. We find the minimums of 0.347 and 0.338 using 6 and 8 variables, respectively.

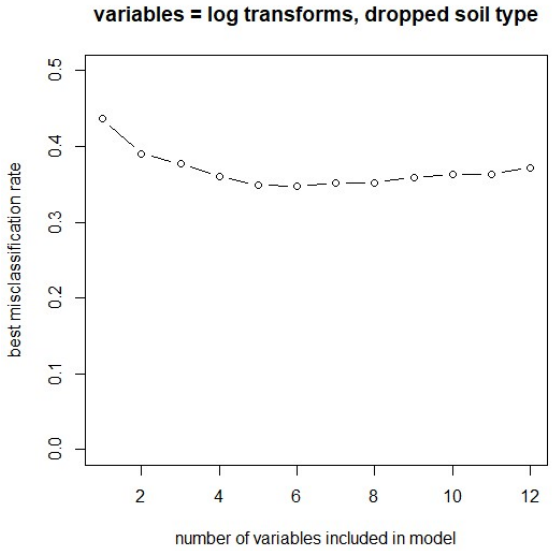

Figure 7: The 5-fold CV misclassification rate for the best model with a given number of variables (shown on the x-axis), generated using RPart. These were selected with all possible log or cube root transformations applied to the dataset, and the transformed soil type variable.

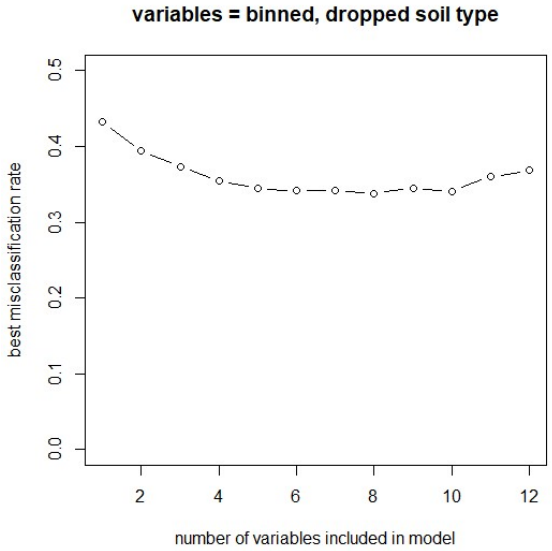

Figure 8: Figure 7: The 5-fold CV misclassification rate for the best model with a given number of variables (shown on the x-axis), generated using RPart. These were selected with all possible binning transformations applied to the dataset, and the transformed soil type variable.

The best performing classification tree used the binned transformations of the distance measurements, and used the formula:

Cover_Type ~ Elevation + Hillshade_Noon + Hillshade_3pm + Wilderness_Area + Aspect_Binned + Horizontal_Distance_To_Hydrology_Binned + Horizontal_Distance_To_Fire_Points_Binned + Horizontal_Distance_To_Roadways_Binned

We then fitted an rpart model using this formula and 10 000 training observations out of 15 120 total observations. The 50% and 80% classification tables from holdout set evaluation can be found in the appendix, as tables A2 and A3. Additionally, this model generated a five-fold CV misclassification rate of 0.338.

We can conclude from these tables that some categories align themselves in pairs – Spruce/Fir and Lodgepole pine, Ponderosa and Douglas fir, and Lodgepole pine and Aspen. These pairs represent categories that are difficult to distinguish from each other. Conversely, Cottonwood/Willow and

Krumholz have misclassification rates under 10% and are easy to distinguish. We would expect this given that Cottonwood/Willow is the only decidouous species in the set, and Krummholz is very unique in that they prefer high altitude environments.

Figure 9 depicts the classification tree for the best performing model. Examining the classification tree allows us to interpret the classes and the explanatory variables most associated with them. Spruce/fir is identified by occupying an elevation between 3038 m and 3233 m. Lodgepole pines are the furthest from roadways. If a tree grows on a south facing slope and it's at between 2370 and 2679 meters elevation, it's likely a ponderosa pine. If it's south facing and below 2370 m, it's likely a Cottonwood/Willow. Aspen are identified by being close to roadways. Douglas firs have northern aspect, and Krummholz grow at elevations greater than 3233 m.

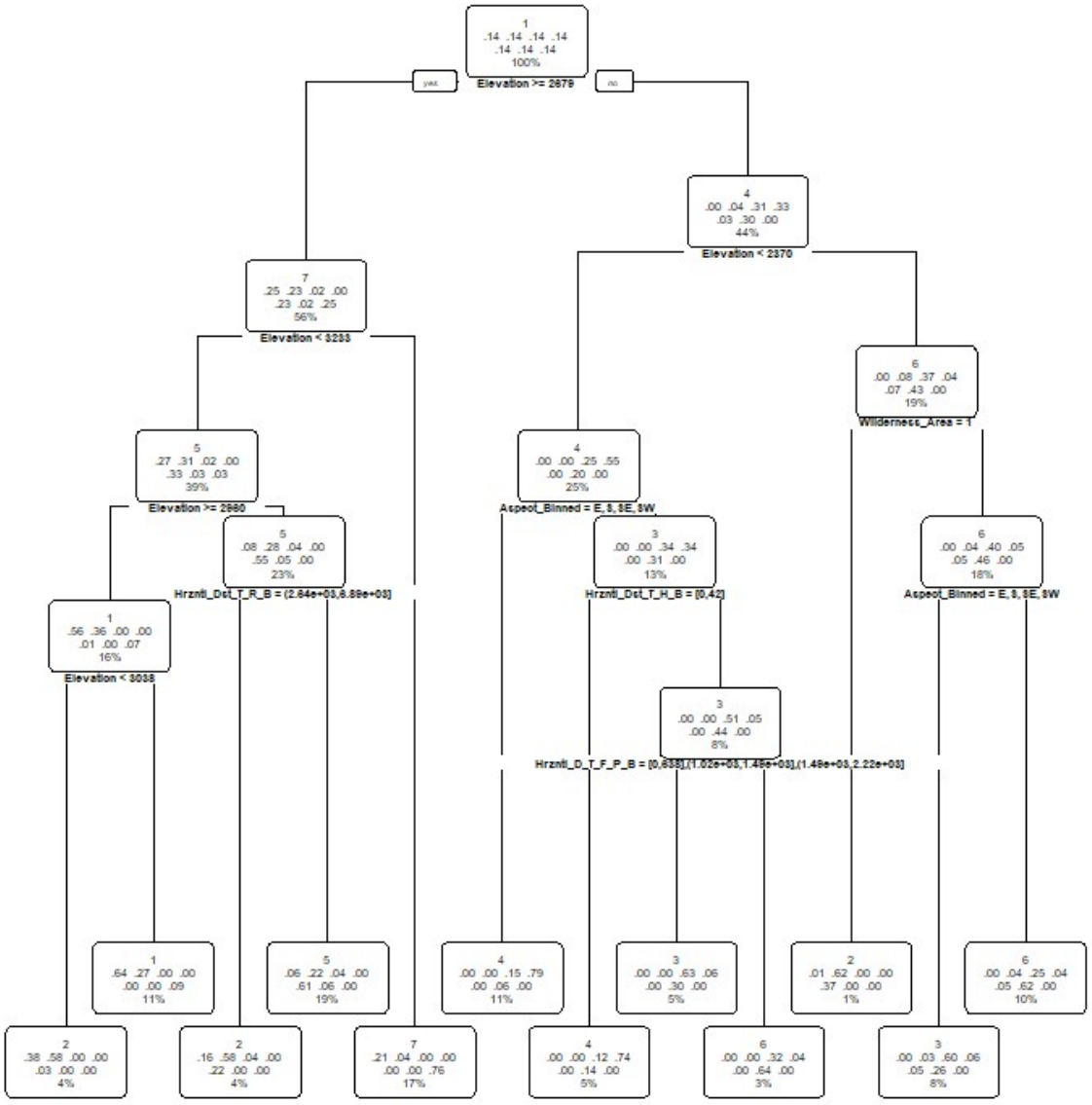

Figure 9: Visual depiction of the best possible classification tree model, generated using RPart.

Random Forest

We applied Random Forests to potentially lower the misclassification error and to select a subset of variables that will lead to better per class predictions from the dataset. Using the dataset that contained both transformed and untransformed variables, we fit a full model containing both transformed and untransformed variables, and then the variables were then selected based on the variable importance criteria, as shown in Figure 10. Random forests can be viewed as a bagging algorithm where only a certain number of variables are considered as split candidates, a general rule of thumb is to use p = $\sqrt{n}$, where n is the number of variables in the model. Such an approach generally helps to decorrelate the trees and allows one to form ensembles of trees that have different significant variables, therefore reducing the effect of correlation between the variables used. As a general rule, once the number of trees in random forest reached about 300, the error rate stabilized. Values up to 1000 were used, but no significant improvement in performance was noted.

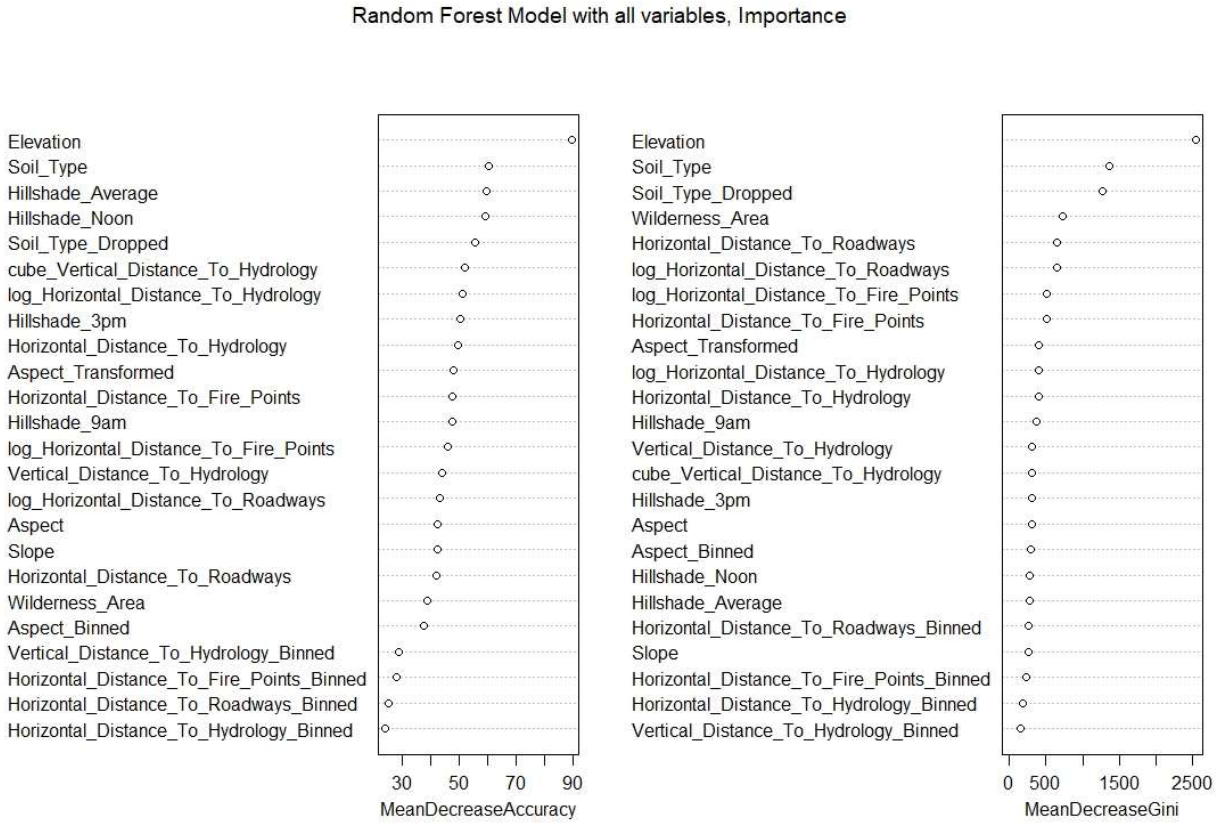

Figure 10: Variable importance plots generated by fitting a random forest model to the dataset with all original variables, and all described transformations.

We compared the performance of models using untransformed variables only (model A) and then a combination of transformed and untransformed, but with either untransformed or transformed version of a given variable was used (model B). Then a subset of top 5 (model C) and top 10 (model D) variables was used to find a more plausible model. These models are described fully in Table 6. A subset of 10 best gave better 5 fold CV performance based on the mean decrease in Gini Index. The full model showed slightly poorer results due to correlation effects. There is a slight variation in the 80% prediction

interval between Spruce / Fir and Lodgepole Pine with Model C producing smaller interval width; however, the misclassification rates are still lower with model D.

Models and Variables	
Model A, Variables without transforms	Elevation, Soil_Type, Wilderness_Area, Horizontal_Distance_To_Roadways, Horizontal_Distance_To_Fire_Points, Aspect, Hillshade_9am, Hillshade_Noon, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Hillshade_3pm, Slope
Model B Selection – transformed or untransformed, but only transformed or untransformed version of a variable at any given time.	Evelation, Soil_Type, Wilderness_Area, Horizontal_Distance_To_Roadways, Horizontal_Distance_To_Fire_Points, log_Horizontal_Distance_To_Fire_Points, Hillshade_3pm, Slope, log_Horizontal_Distance_To_Hydrology, Aspect_Tansformed, Hillshade_9am, Hillshade_Noon
Model C Top 5, based on mean decrease in Gini index	Elevation + Soil_Type, Soil_Type_Dropped, Wilderness_Area, Horizontal_Distance_To_Roadways
Model D Top 10, based on mean decrease in Gini index	Elevation, Soil_Type, Soil_Type_Dropped, Wilderness_Area, Horizontal_Distance_To_Roadways, log_Horizontal_Distance_To_Roadways, Horizontal_Distance_To_Fire_Points, log_Horizontal_Distance_To_Fire_Points, log_Horizontal_Distance_To_Hydrology, Aspect_Transformed
Model E Full – transformed and untransformed variables	All Variables

Table 6: The variables included in each random forest model measured

	Spruce / Fir	Lodgepole Pine	Ponderosa Pine	Cottonwood / Willow	Aspen	Douglas-Fir	Krummholz	Overall
A	0.234	0.301	0.173	0.028	0.048	0.119	0.032	0.133
B	0.219	0.286	0.163	0.027	0.042	0.110	0.027	0.125
C	0.309	0.377	0.284	0.116	0.091	0.336	0.064	0.226
D	0.221	0.264	0.161	0.029	0.043	0.111	0.027	0.123
E	0.212	0.295	0.167	0.031	0.042	0.107	0.029	0.126

Table 7: Misclassification error for different random forest models and cover types based on 5-fold cross-validation, along with overall CV misclassification rate

	Spruce / Fir	Lodgepole Pine	Ponderosa Pine	Cottonwood / Willow	Aspen	Douglas-Fir	Krummholz
A	1.990	2.094	1.935	1.354	1.743	1.943	1.322
B	1.976	2.083	1.913	1.354	1.687	1.910	1.330
C	1.736	1.861	1.837	1.370	1.542	1.799	1.286
D	1.839	1.948	1.756	1.272	1.500	1.744	1.245
E	2.074	2.162	1.945	1.348	1.810	1.962	1.357

Table 8: 80% Prediction Interval Width, based on 5-fold cross-validation, for each cover type for different random forest models.

	Spruce / Fir	Lodgepole Pine	Ponderosa Pine	Cottonwood / Willow	Aspen	Douglas-Fir	Krummholz
Errors	0.221	0.264	0.161	0.029	0.043	0.111	0.027
50% PI	1.116	1.174	1.117	1.046	1.077	1.110	1.024
80% PI	1.839	1.948	1.756	1.272	1.500	1.744	1.245

Table 9: Model D CV misclassification rates, and 50% and 80% prediction interval widths, for each cover type.

Model D produced the best results, as shown in tables 7 and 8. However, no combination of variables have improved separation between different classes. Model performance and importance suggest more important variables as well: Elevation, Soil Type, Wilderness Area, and the distance measurements were found to be important variables, and hill shade index and slope were less important. Notably, Spruce/Fir, Lodgepole Pine, Ponderosa Pine, and Cottonwood/Willow displayed the same misclassification patterns as the top logistic regression models and gradient boosting model.

Model Comparison and Interpretations

As was expected, random forest and gradient boosting models performed the best, evaluated by CV misclassification rate. The two top performing models of these, based on overall CV error, were Model D for random forest (Top ten variables selected based on Gini index) and the final best-performing gradient boosting model. Of these two, the random forest best performer had a marginally better overall CV error rate of less than a percent. However, the random forest model had larger average 80% and 50% prediction interval lengths for all classes. As such, with only a marginal improvement in CV error for the random forest, we can conclude that the best gradient boosting model is the best predictor for all models attempted based on overall prediction error.

However, this is under the assumption that the real world contains a similar proportion of each cover type, as the dataset used in this report contained equal numbers of each class. This is likely not the case. For all models attempted, there was variance in misclassification rates per class, with Spruce/Fir, Lodgepole Pine, Douglas Fir, and Ponderosa Pine being misclassified more often than the other classes. Our prediction methods will perform poorer if these classes are overrepresented.

If we compare the top performing random forest model to the top performing gradient boosting model, we see that in these “problem” classes misclassification rates are more than a percent lower in the random forest model for Spruce/Fir and Lodgepole pine, and similar if not better for all other classes. As such, if we expect that the class proportions are not equal in a prediction setting, despite the larger prediction interval widths, it would be better to use the random forest model.

One of the goals for this analysis is to is to contribute to developing the niche space, or set of favourable environmental conditions, for each cover type. The four “problem” classes display a consistent pattern of misclassification across all models, with Spruce/Fir and Lodgepole Pine often misclassified as each other, and Ponderosa Pine and Douglas Fir often misclassified as each other. This indicates that, using the variables in this dataset, these pairs of trees have similar if not overlapping niche spaces.

With regards to important predictors, different variables with different transformations were important for prediction in different models attempted. However, in general, Elevation, Soil Type, and the four distance measurements (ex: Horizontal Distance to Hydrology) appeared to be the most important variables (in one transformation or another) across all models attempted, for prediction. This is indicated in the variable importance plots for random forest and gradient boosting, or in the best

performing models after variable selection for logistic regression and classification trees. Across all models, Elevation appeared to be the most important prediction variable, with high values in variable importance plots and appearing in all logistic regression and classification tree models. This implies that, of the environmental variables studied, these trees overlap the least with regards to elevation, the type of soil, and their distance from meaningful environmental features.

Unfortunately, our best predicting models are not suitable for variable interpretation, so this task must involve the results from the poorer performing classification tree and logistic regression models. Additionally, interpretations will not be given for the soil type, as with 20 different levels (after transformation) the results are too detailed for this report. However, with the remaining important variables we can provide the following interpretations for each class:

- **Spruce/Fir:** Similar in niche space to Lodgepole pine. Generally found in higher altitudes than the other cover types examined, except for Krummholz, and generally found closer horizontally to hydrology than most other cover types. Roughly in the middle of the classes studied for distances to roadways and fire points.
- **Lodgepole Pine:** Similar in niche space to Spruce/Fir. Generally found in higher altitudes than other cover types, but lower than Spruce/Fir and Krummholz. Horizontally farther than all classes except Ponderosa to hydrology. Roughly in the middle of classes studied for distances to roadways and fire points.
- **Ponderosa Pine:** Similar in niche space to Douglas-Fir. Occupies the second-lowest elevation of all cover types and is horizontally the farthest from hydrology of all classes. It is the closest of all classes to fire points, but roughly in the middle of classes for distance to roadways.
- **Cottonwood/Willow:** Found the closest horizontally to hydrology, as is expected. Found the second farthest from fire points, and the farthest from roadways of all classes, and is found at the lowest elevations of all studied.
- **Aspen:** Roughly in the middle of all cover types studied for elevation and horizontal distance to hydrology. Generally, the second closest of cover types to fire points, and the closest to roadways.
- **Douglas-Fir:** Similar in niche space to Ponderosa Pine. Occupies a higher elevation than Ponderosa pine in general and found farther from hydrology. Roughly in the middle of cover types for distance to roadways and fire points.
- **Krummholz:** Found at the highest elevations and the farthest from fire points, but the second closest to roadways and hydrology.

For more detailed delineations between classes for certain variables, see the classification tree in the Figure 9. Additionally, the vertical distance to hydrology variable was found to be strongly correlated with elevation. As such, the interpretations from the coefficients in the logistic regression model appeared nonsensical at times (ex: Cottonwood/Willow is close to hydrology horizontally, but far vertically). As such, this variable was not included in interpretations, even though it improved prediction error.

Contributions

Armandas: Classification tree model
Bryn: Multinomial Logistic Regression model and Gradient Boosting model
Dmitri: Random Forest model

Appendix:

Soil Types:

The soil types are:

1 Cathedral family - Rock outcrop complex, extremely stony.
2 Vanet - Ratake families complex, very stony.
3 Haploborolis - Rock outcrop complex, rubbly.
4 Ratake family - Rock outcrop complex, rubbly.
5 Vanet family - Rock outcrop complex complex, rubbly.
6 Vanet - Wetmore families - Rock outcrop complex, stony.
7 Gothic family.
8 Supervisor - Limber families complex.
9 Troutville family, very stony.
10 Bullwark - Catamount families - Rock outcrop complex, rubbly.
11 Bullwark - Catamount families - Rock land complex, rubbly.
12 Legault family - Rock land complex, stony.
13 Catamount family - Rock land - Bullwark family complex, rubbly.
14 Pachic Argiborolis - Aquolis complex.
15 unspecified in the USFS Soil and ELU Survey.
16 Cryaquolis - Cryoborolis complex.
17 Gateview family - Cryaquolis complex.
18 Rogert family, very stony.
19 Typic Cryaquolis - Borohemists complex.
20 Typic Cryaquepts - Typic Cryaquolls complex.
21 Typic Cryaquolls - Leighcan family, till substratum complex.
22 Leighcan family, till substratum, extremely bouldery.
23 Leighcan family, till substratum - Typic Cryaquolls complex.
24 Leighcan family, extremely stony.
25 Leighcan family, warm, extremely stony.
26 Granile - Catamount families complex, very stony.
27 Leighcan family, warm - Rock outcrop complex, extremely stony.
28 Leighcan family - Rock outcrop complex, extremely stony.
29 Como - Legault families complex, extremely stony.
30 Como family - Rock land - Legault family complex, extremely stony.
31 Leighcan - Catamount families complex, extremely stony.
32 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
33 Leighcan - Catamount families - Rock outcrop complex, extremely stony.
34 Cryorthents - Rock land complex, extremely stony.
35 Cryumbrepts - Rock outcrop - Cryaquepts complex.
36 Bross family - Rock land - Cryumbrepts complex, extremely stony.
37 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
38 Leighcan - Moran families - Cryaquolls complex, extremely stony.
39 Moran family - Cryorthents - Leighcan family complex, extremely stony.
40 Moran family - Cryorthents - Rock land complex, extremely stony.

Table A1: 50% Prediction interval frequencies for the "best" logistic regression model.

Actual	**1**	**12**	**15**	**17**	**2**	**21**	**25**	**26**	**3**
1	**652**	**29**	**6**	**11**	**145**	**25**	**10**	**0**	**0**
2	**241**	**36**	**9**	**3**	**474**	**48**	**46**	**2**	**15**
3	**0**	**0**	**0**	**0**	**5**	**0**	**3**	**1**	**583**
4	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**45**
5	**9**	**9**	**2**	**0**	**119**	**11**	**37**	**6**	**14**
6	**0**	**0**	**0**	**0**	**6**	**0**	**7**	**2**	**270**
7	**86**	**6**	**0**	**25**	**1**	**0**	**0**	**0**	**0**

Actual	**32**	**34**	**35**	**36**	**4**	**43**	**46**	**5**	**51**
1	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**33**	**6**
2	**0**	**0**	**2**	**5**	**0**	**0**	**0**	**148**	**7**
3	**3**	**11**	**12**	**33**	**86**	**13**	**4**	**22**	**0**
4	**0**	**8**	**0**	**15**	**946**	**13**	**9**	**0**	**0**
5	**0**	**0**	**0**	**3**	**0**	**0**	**0**	**795**	**0**
6	**0**	**5**	**3**	**21**	**33**	**4**	**1**	**9**	**0**
7	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Actual	**52**	**53**	**56**	**57**	**6**	**62**	**63**	**64**	**65**	**7**	**71**	**75**
1	**8**	**0**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	**116**	**8**	**0**
2	**25**	**3**	**3**	**0**	**33**	**4**	**2**	**0**	**6**	**14**	**1**	**1**
3	**4**	**14**	**4**	**0**	**186**	**0**	**35**	**0**	**5**	**0**	**0**	**0**
4	**0**	**0**	**0**	**0**	**23**	**1**	**5**	**7**	**0**	**0**	**0**	**0**
5	**43**	**6**	**10**	**0**	**17**	**2**	**7**	**0**	**8**	**1**	**0**	**0**
6	**5**	**1**	**7**	**0**	**660**	**4**	**35**	**5**	**19**	**0**	**0**	**0**
7	**0**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**953**	**12**	**0**

Table A2: The 50% Prediction Interval Frequencies for the top performing classification tree model

actual	1	2	3	4	5	6	7
1	932	309	3	0	157	4	468
2	404	737	29	1	553	54	86
3	0	22	990	306	11	458	0
4	0	0	101	1753	0	60	0
5	0	191	47	0	1574	62	0
6	0	0	440	183	173	1056	0
7	144	0	0	0	0	4	1708

Table A3: The 80% Prediction Interval Frequencies for the top performing classification tree model

Actual	12	21	25	251	36	43	52	63	71
1	932	223	2	84	3	0	157	4	468
2	404	337	105	295	29	1	553	54	86

3	0	0	0	22	990	306	111	458	0
4	0	0	0	0	101	1753	0	60	0
5	0	17	63	111	47	0	1574	62	0
6	0	0	0	0	440	183	173	1056	0
7	144	0	0	0	0	0	4	0	1708