Predicting Forest Cover Type

Team Checkpoint 1: Modeling Problem, Data Quality & EDA

Dong Bing, Richard Gunawarden, Kent Merdes, Christina Macholan, Tyler Wintermeyer

1. Introduction

Having a complete and accurate record of natural resources helps local and national resource management organizations make informed decisions about how to best preserve or utilize specific types of land. In this study, we examine whether predictive models that employ digital spatial data from geographic information system (GIS) can substitute manual field surveys in order to correctly categorize land forest cover type.

According to Blackard and Dean's original paper on this subject, "generally, forest cover type data is either directly recorded by field personnel or estimated from remotely sensed data"[include citation and reference], which can be time-consuming and expensive work. Our goal is to assess whether this time and cost could be reduced by relying on data available through digital collection, instead.

[Placeholder for literature review section about types of models used in our project.]

2. The Modeling Problem

Our aim is to build a multi-class classification model that can predict the forest cover type for a 30 x 30 meter parcel of land. To do so, we will use data obtained from the US Forest Service (USFS) Region 2 Resource Information System (RIS) which contains measured attributes for over half a million parcels of land. These attributes include wilderness area type, soil type, the parcel's orientation to water, roadways, and wildfire-prone areas, and other cartographic features.

3. The Data

The data used for our study comes from a 1998 study by Blackard and Dean in which US Forest Service (USFS) Region 2 Resource Information System (RIS) data and US Geological Survey (USGS) were compiled for 581,012 parcels of land. Each parcel corresponds with a 30 x 30 meter area in one of the following four regions of the Roosevelt National Forest in Colorado: Rawah, Neota, Comanche Peak, Cache la Poudre. These regions are mapped out in Figure 1.

Figure 1: Map of wilderness areas included in study from Blackard and Dean's original paper

The 13 variables listed in Table 1 were provided for each 30 x 30 meter land parcel (equivalent to one observation in the dataset). For prediction purposes, Forest Cover Type will be used as the dependent variable, and combinations of some or all of the remaining variables as the predictor variables.

Table 1: Description of variables

Variable Description	Type & Measurement
Elevation	numeric in meters
Aspect	numeric in degrees azimuth
Slope	numeric in degrees
Horizontal distance to nearest source water features	numeric in meters
Vertical distance to nearest source water features	numeric in meters
Horizontal distance to nearest roadway	numeric in meters
Hillshade index at 9am during summer solstice	numeric as index (0 to 255)
Hillshade index at Noon during summer solstice	numeric as index (0 to 255)
Hillshade index at 3pm during summer solstice	numeric as index (0 to 255)
Horizontal distance to nearest wildfire ignition points	numeric in meters
Wilderness Areas (4 areas)	binary for each area (0 or 1)
- 1 - Rawah	
- 2 - Neota	
- 3 - Comanche Peak	
- 4 - Cache la Poudre	
Soil Type (40 types)	binary for each type (0 or 1)
- See Appendix A for details	
Forest Cover Type (7 types)	integer for each type (0 or 1)
- 1 - Spruce/Fir	
- 2 - Lodgepole Pine	
- 3 - Ponderosa Pine	
- 4 - Cottonwood/Willow	
- 5 - Aspen	
- 6 - Douglas-fir	
- 7 - Krummholz	

4. Exploratory Data Analysis (EDA)

Traditional EDA

An initial examination of the forest cover types in the provided data set shows that the most common types of cover are, by far, Lodgepole Pine (49% of records) and Spruce/Fir (36% of records). Other tree cover types are much rarer for the areas surveyed (<10% of records each) and therefore need to be oversampled in the training data set for building models using Artificial Neural Networks. Figure 2 shows a breakdown of the frequency of each cover type from the full data set.

To check the data quality, we first reviewed summary statistical measurements for each variable. None of the variables have missing values, and the range of values for each metric seems reasonable (no unexpected negative or zero-value measurements).

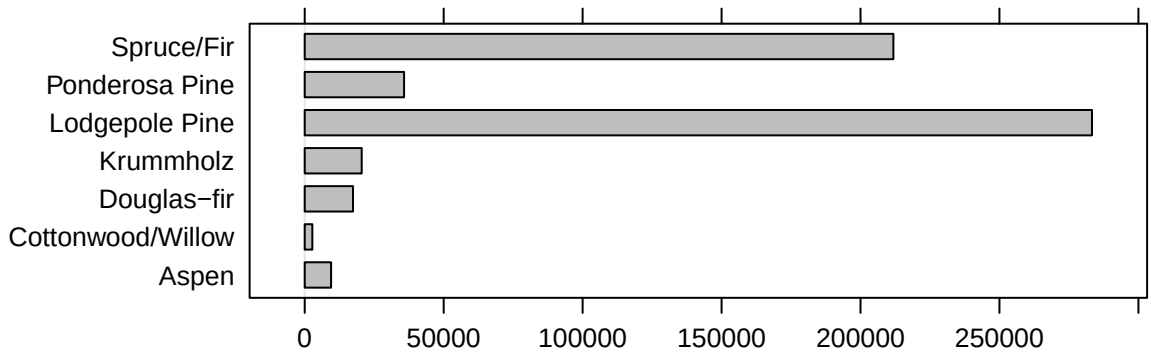

Figure 2: Forest Cover Type Frequency

Numeric variables

To check for possible outliers and data distribution, we created density plots and boxplots broken out by forest cover class for each numeric variable, as shown in Figures 3 and 4. Elevation has a relatively normal distribution across all tree types, whereas other variables do not. The right-skewed variables (Horizontal Distance to Hydrology, Vertical Distance to Hydrology, and Horizontal Distance to Firepoints) and the left-skewed variables (Hillshade at 9am and Hillshade at Noon) may need to undergo transformations for any modeling procedures that assume normality for the predictor variables. Aspect is a unique variable in that it shows a bimodal distribution.

Elevation appears to be the most differentiating numeric variable across forest cover types, which makes it an especially good candidate for inclusion in our models.

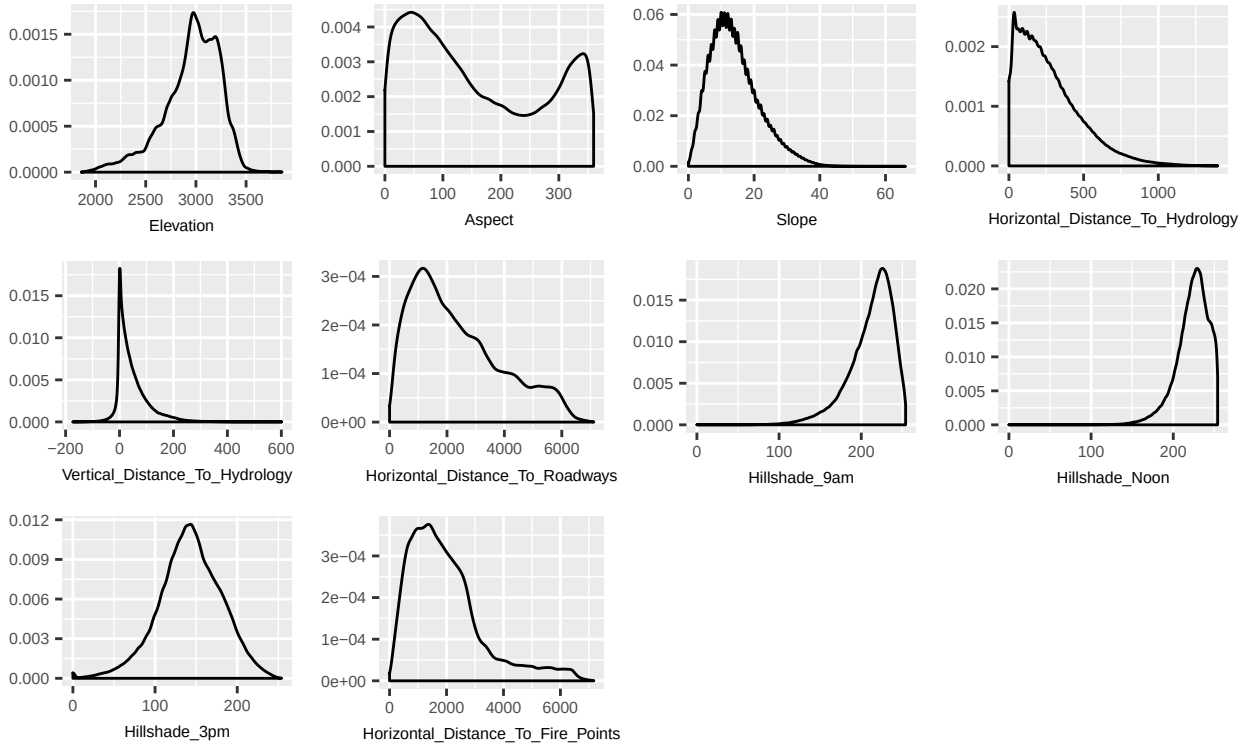

Figure 3: Density plots of numeric variables

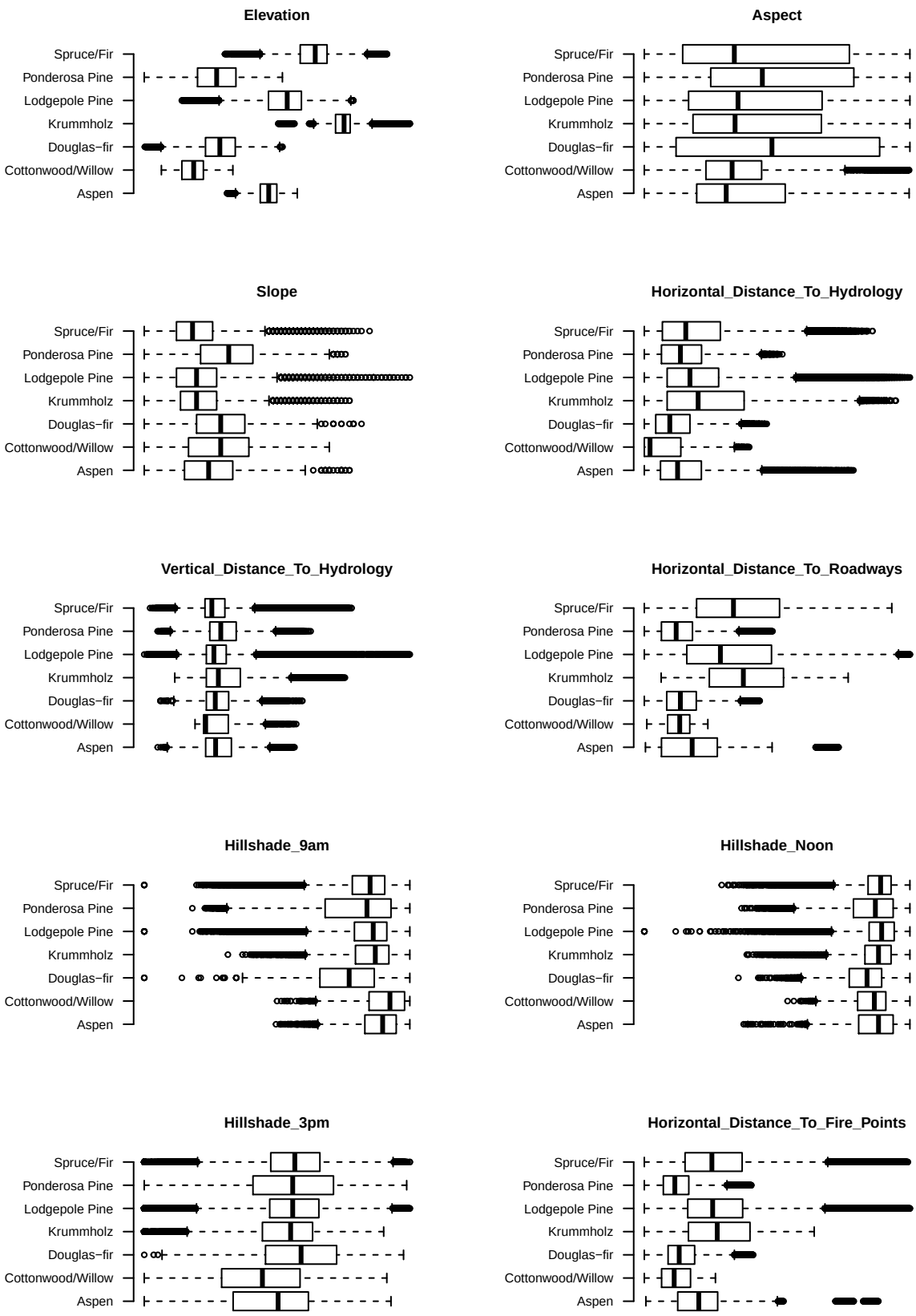

Figure 4: Boxplots of numeric variables by Forest Cover Type

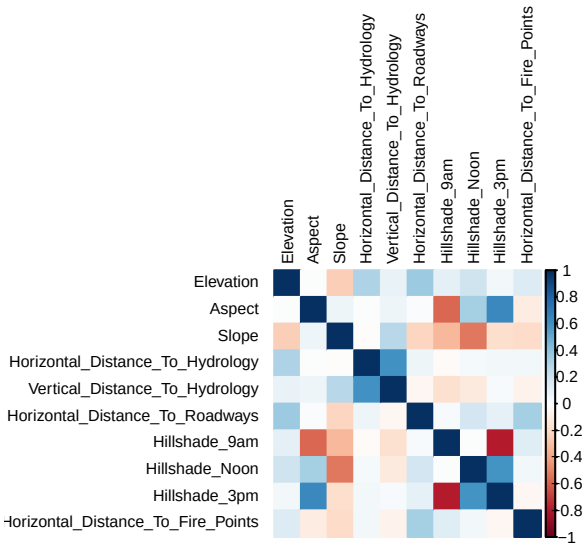

Figure 5: Correlations for Numeric Variables

To understand the relationships between the numeric predictor variables, we can examine the correlation plot in Figure 5.

There are six pairwise correlations that have a value higher than absolute 0.5

- Hillshade at 9am, Hillshade at 3pm (-0.78)
- Aspect, Hillshade at 3pm (0.65)
- Horizontal Distance to Hydrology, Vertical Distance to Hydrology (0.61)
- Slope, Hillshade at Noon (-0.61)
- Hillshade at Noon, Hillshade at 3pm (0.59)
- Aspect, Hillshade at 9am (-0.58)

The scatterplots in Figure 6 help us examine these highly correlated variables more closely. From the plots, we observe the following: * The hillshade at noon and 3pm creates an ellipsoid pattern. * As the horizontal distance to a hydrology increases, the variance in vertical distance to hydrology increases. * As slope increases hillshade at noon decreases, with wider variance at steeper slopes. * Hillshade at 3pm has a sigmoidal relationship with Aspect. * Aspect and Hillshade at 9am have a more defined sigmoidal relationship.

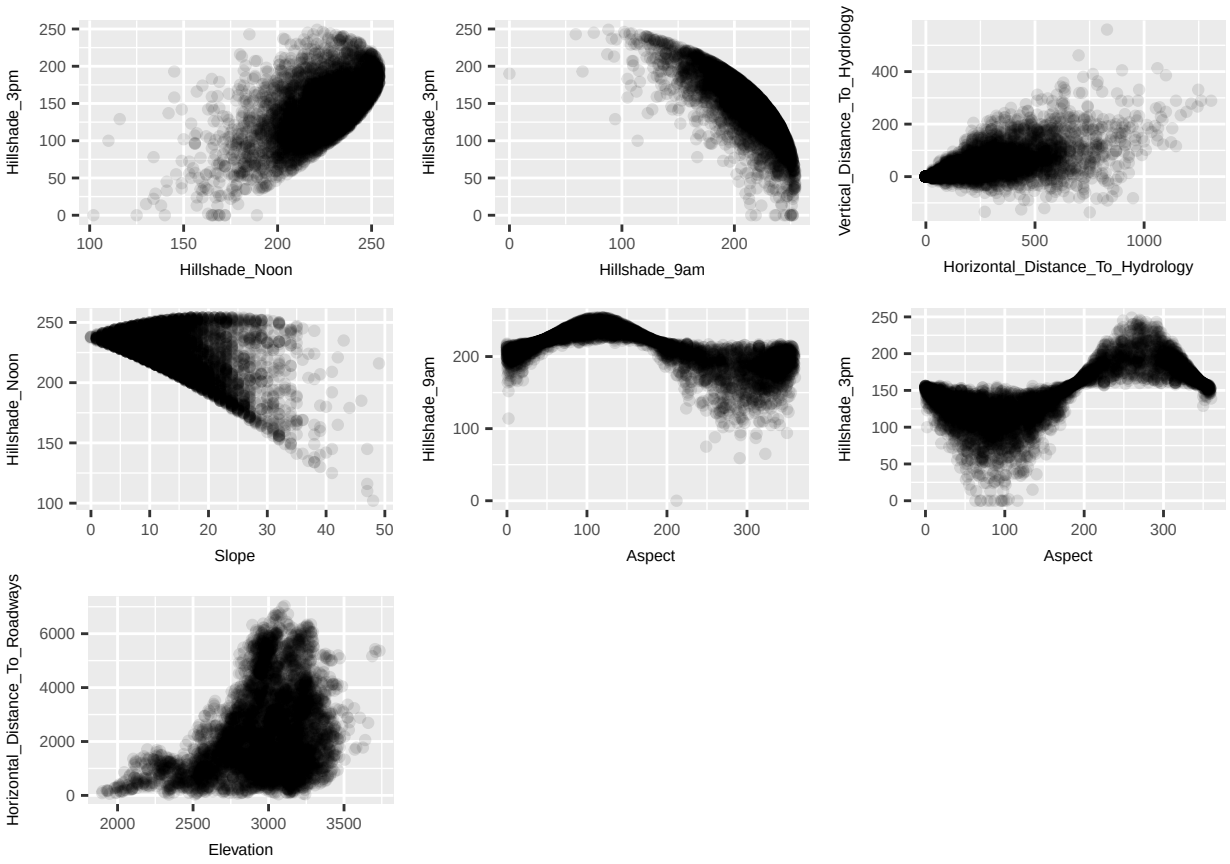

Figure 6: Scatterplots of Highly Correlated Numeric Variables

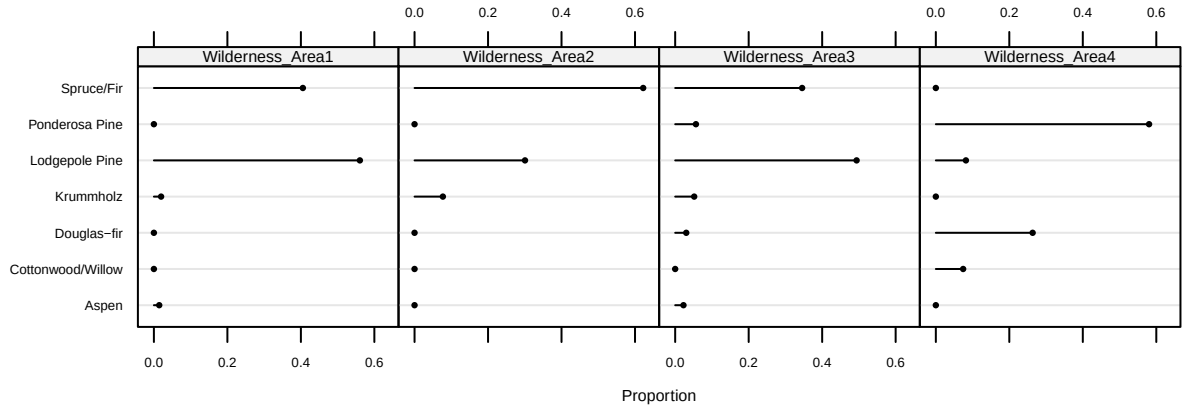

Figure 7: Proportion of Cover Type within each Wilderness Area

Categorical variables

A dot plot of forest cover type by wilderness area in Figure 7 shows the following trends:

- Wilderness Area 1 has a higher proportion of Lodgepole Pines than any of the other areas.
- Wilderness Area 2 has a higher proportion of Spruce/Fir and Krumholz trees than any of the other areas.
- Wilderness Area 3 has a higher proportion of Aspens than any of the other areas.
- Wilderness Area 4 has a higher proportion of Ponderosa Pines, Douglas Firs, and Cottonwood/Willow trees than any of the other areas.

A dot plot of forest cover type by soil type in Figure 8 also shows variability in the proportion of tree types from soil to soil. For example, some soil types show a very high proportion of Lodgepole Pines (e.g. Soil Type 7) whereas others show a very low proportion of Lodgepole Pines (e.g. Soil Type 37). The distinctive make-up of trees by soil type suggests that this may be a good predictive factor for our models.

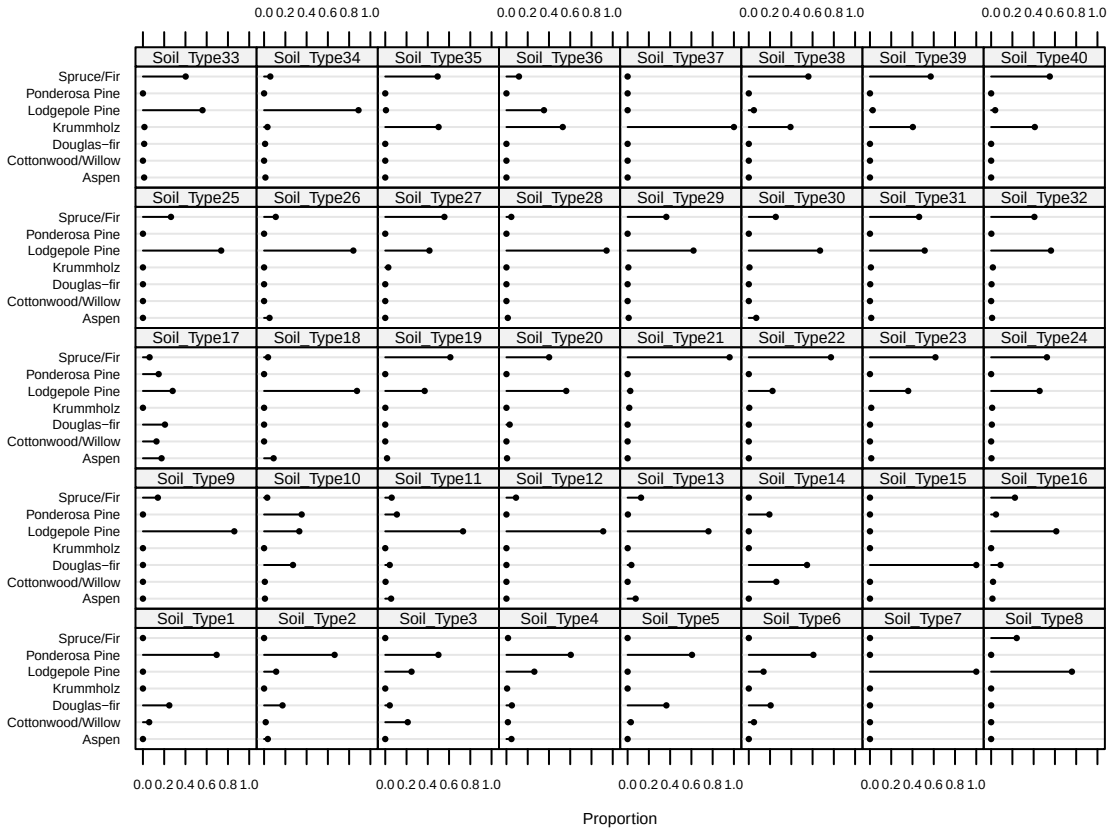

Figure 8: Proportion of Forest Cover by Soil Type

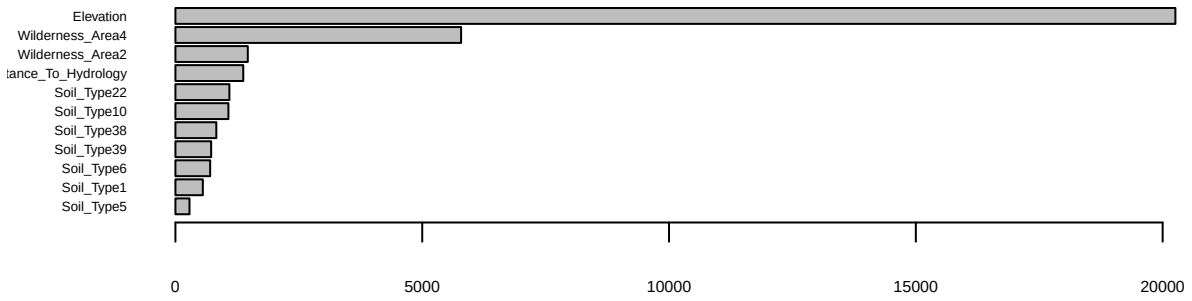

Figure 9: Naive Classification Tree Variable Importance Plot

Model-Based EDA

Model-based exploratory data analysis allows us to glean additional information about the relationships between our predictors and the response variable from naive models applied to the training dataset. In particular, tree-based classification can reveal which features are most important for prediction and any possible interaction effects that are not initially apparent from univariate and bivariate exploratory data analysis.

Simple Tree-based Classification Model

By creating a simple tree-based classification model, we confirm that Elevation is the most important predictor of forest cover type.

Other variables that could be important to defining models are shown in the variable importance plot in Figure 9.

A plot of the decision tree in Figure 10 shows that this model has significant limitations and ignores the rarer tree types in partitioning the data. With an accuracy of only 7%, we learn that a much more sophisticated model will be required to predict cover type.

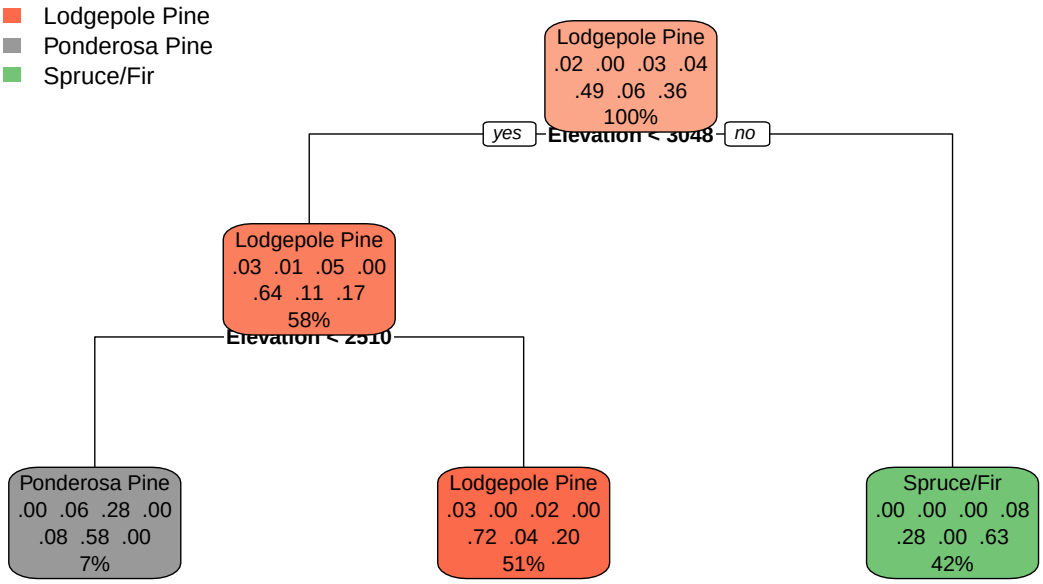

Figure 10: Naive Classification Tree

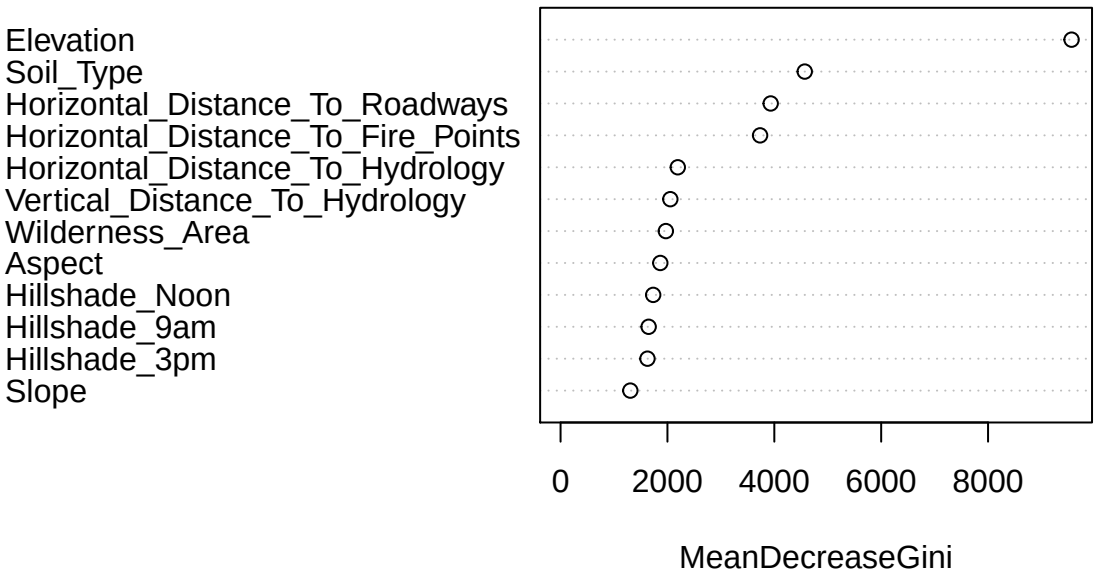

Figure 11: Random Forest Variable Importance Plot

Random Forest Model

According to the importance plot for a random forest model, the elevation variable has the higest importance, followed by soil type, the horizontal distance to roadways, and the horizontal distance to wildfire points. Figure 11 shows the ranking and relative importance of the variables selected.

Boruta Model

According to the boruta algorithm for feature selection, elevation, again, has the highest importance. *[Add comments on variable importance here after running algorithm – couldn't get results on Christina's machine.]*

Next Steps for our Paper

- Reach final agreement on sampling approach (i.e. sample randomly across the full dataset vs. sample randomly within each forest cover type category).
- Begin building models four to five possible models using Artificial Neural Networks, SVMs, Lasso, and Ridge Regression.
- Figure out how to properly do citations in R Markdown.
- Continue refining introduction to align with results from the modeling process.
- Begin writing for additional sections of the paper.
- Continue to revise EDA – include only what's most important once modeling is complete.

5. Predictive Modeling: Methods and Results

Train / Test Data

Individual Model A

Individual Model B

6. Comparison of Results

7. Conclusions

8. Bibliography

9. Appendices

Appendix A: Data Keys

Descriptions of 40 Soil Types

Study	Code	USFS ELU Code Description
1	2702	Cathedral family - Rock outcrop complex, extremely stony.
2	2703	Vanet - Ratake families complex, very stony.
3	2704	Haploborolis - Rock outcrop complex, rubbly.
4	2705	Ratake family - Rock outcrop complex, rubbly.
5	2706	Vanet family - Rock outcrop complex complex, rubbly.
6	2717	Vanet - Wetmore families - Rock outcrop complex, stony.
7	3501	Gothic family.
8	3502	Supervisor - Limber families complex.
9	4201	Troutville family, very stony.
10	4703	Bullwark - Catamount families - Rock outcrop complex, rubbly.
11	4704	Bullwark - Catamount families - Rock land complex, rubbly.
12	4744	Legault family - Rock land complex, stony.
13	4758	Catamount family - Rock land - Bullwark family complex, rubbly.
14	5101	Pachic Argiborolis - Aquolis complex.
15	5151	unspecified in the USFS Soil and ELU Survey.
16	6101	Cryaquolis - Cryoborolis complex.
17	6102	Gateview family - Cryaquolis complex.
18	6731	Rogert family, very stony.
19	7101	Typic Cryaquolis - Borohemists complex.
20	7102	Typic Cryaquepts - Typic Cryaquolls complex.
21	7103	Typic Cryaquolls - Leighcan family, till substratum complex.
22	7201	Leighcan family, till substratum, extremely bouldery.
23	7202	Leighcan family, till substratum - Typic Cryaquolls complex.
24	7700	Leighcan family, extremely stony.
25	7701	Leighcan family, warm, extremely stony.
26	7702	Granile - Catamount families complex, very stony.
27	7709	Leighcan family, warm - Rock outcrop complex, extremely stony.
28	7710	Leighcan family - Rock outcrop complex, extremely stony.
29	7745	Como - Legault families complex, extremely stony.
30	7746	Como family - Rock land - Legault family complex, extremely stony
31	7755	Leighcan - Catamount families complex, extremely stony.
32	7756	Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
33	7757	Leighcan - Catamount families - Rock outcrop complex, extremely stony.
34	7790	Cryorthents - Rock land complex, extremely stony.
35	8703	Cryumbrepts - Rock outcrop - Cryaquepts complex.
36	8707	Bross family - Rock land - Cryumbrepts complex, extremely stony.
37	8708	Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
38	8771	Leighcan - Moran families - Cryaquolls complex, extremely stony.
39	8772	Moran family - Cryorthents - Leighcan family complex, extremely stony.
40	8776	Moran family - Cryorthents - Rock land complex, extremely stony.

Lookup table for Soil Code

Climatic Zone (first digit)	Geologic Zones (second digit)
1. lower montane dry	1. alluvium
2. lower montane	2. glacial
3. montane dry	3. shale
4. montane	4. sandstone
5. montane dry and montane	5. mixed sedimentary
6. montane and subalpine	6. unspecified in the USFS ELU Survey
7. subalpine	7. igneous and metamorphic
8. alpine	8. volcanic

Appendix B: Additional EDA