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 spacial 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 multiclass calssification 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.

For each parcel of land, the 13 variables listed in Table 1 were provided.

*** Table 1: Description of variables

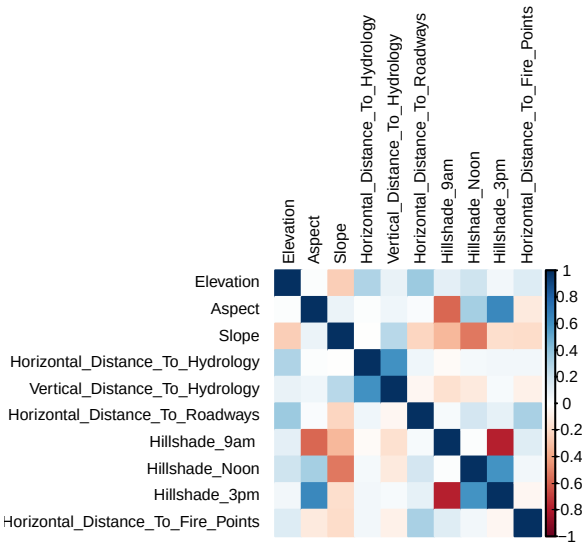
Variable Description	**Type & Measurement**
Elevation	numeric in meters
Aspect	numeric in degrees azimuth
Slope	numeric in degrees
Horizontal distance to nearest surce water features	numeric in meters
Vertical distance to nearest surce water features	numeric in meters
Horizontal distance to nearest roadway	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)

Variable Description	Type & Measurement
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

Figure 3: Correlation plot of numeric variables



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 any possible interaction effects that are not initially apparent from univariate and bivariate exploratory data analysis.

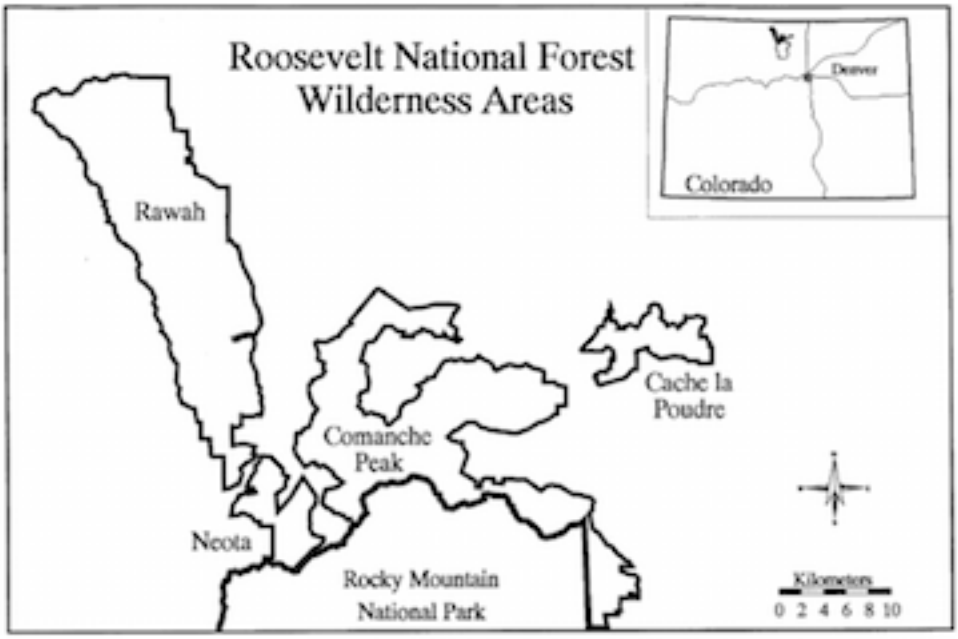


Figure 1: Map of wilderness areas included in study from Blackard and Dean's original paper

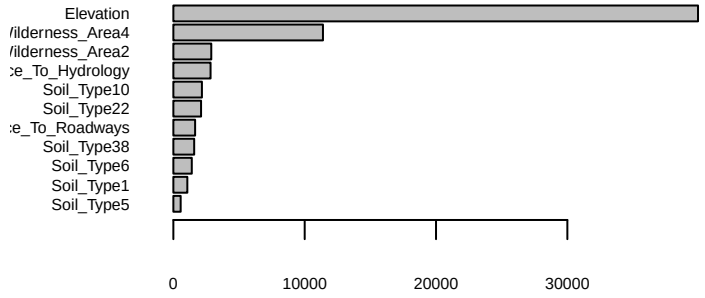
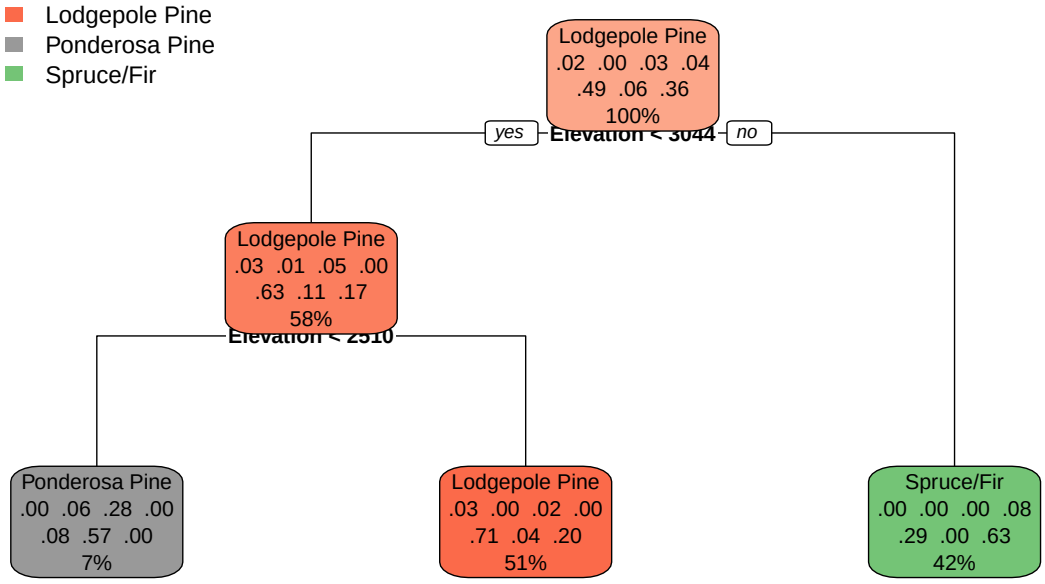


Figure 2: Classification Tree Variable Importance Plot

Tree-based Classification Model

Figure 6: Classification tree



K-means Clustering Placeholder

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 using Neural Networks and SVMs
- 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

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

```
densityplot(~covtype$Elevation|covtype$Cover_Type, include=FALSE, results=FALSE)
```