Predicting Forest Cover Type

Team Checkpoint 1

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

1. Introduction

- Provide an overview and general statement of the problem.
- Provide a general discussion of how you have approached the problem and highlight some of the interesting results.

According to Blackard and Dean's original paper, "forest cover type data is either directly recorded by field personnel or estimated from remotely sensed data" [see @blackarddean, p. 132].

Four wilderness areas in Roosevelt National Forest: Rawah, Neota, Comanche Peak, Cache la Poudre.

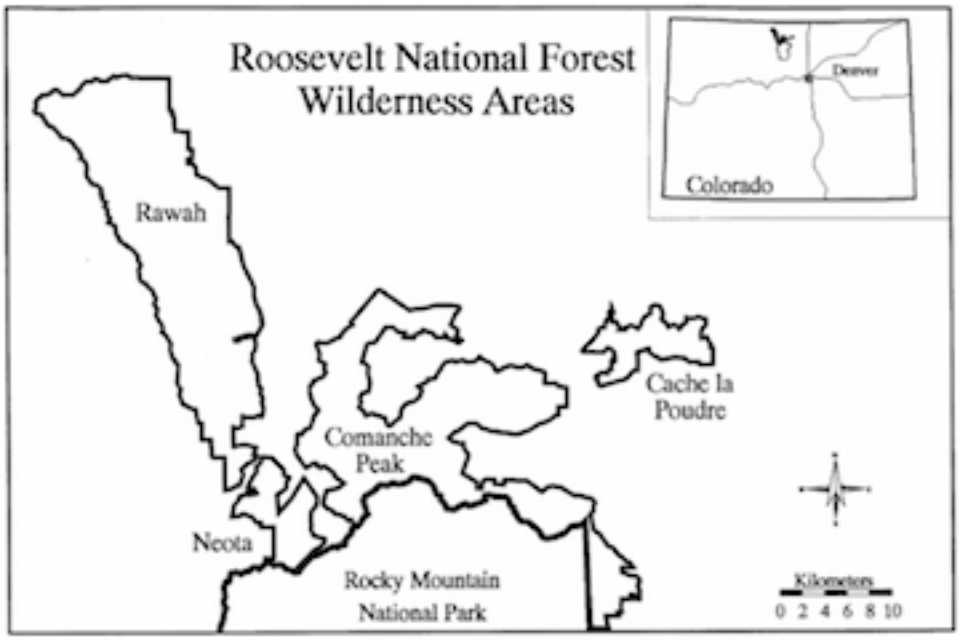


Figure 1: Map of wilderness areas included in study from Blackard and Dean's original paper

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, topographical features, and the parcel's orientation to water, roadways, and wildfire-prone areas.

3. The Data

581,012 records 55 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)
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*	

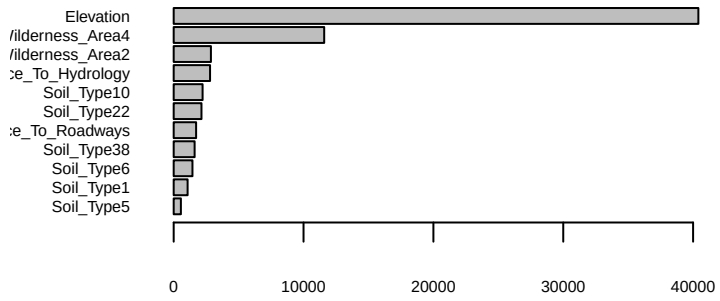
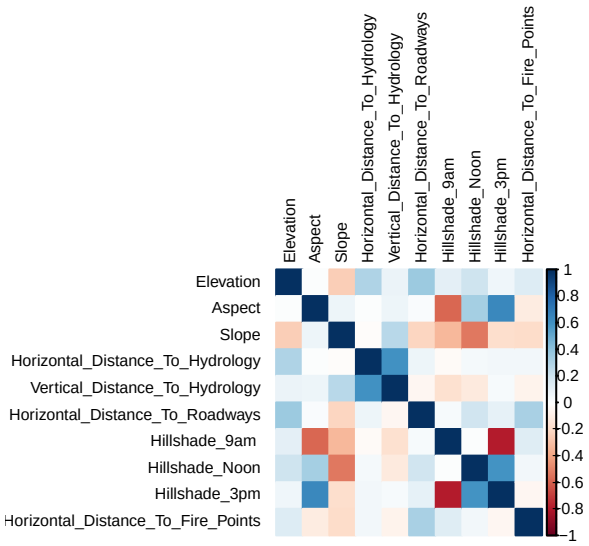


Figure 2: Classification Tree Variable Importance Plot

4. Exploratory Data Analysis (EDA)

Traditional EDA

Figure 3: Correlation plot of numeric variables

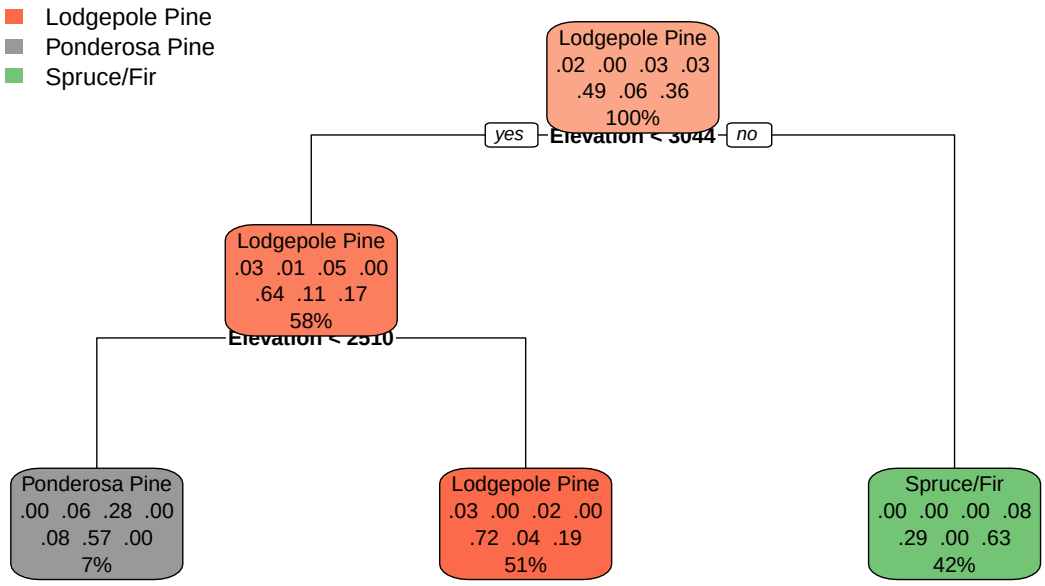


Model-Based EDA

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.

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

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