Determining Forest Cover Type Using Classification

Francesca Chiappetta

Intro

Background

- Growing demand of carbon credits is leading to a burgeoning carbon market.

- One way to generate carbon credits is through sequestering carbon in forest ecosystems.

Context

- Determining the amount of carbon sequestered by a tree or many tree is complicated.

- It heavily depends on type of tree (and DBH)

- If this data is available, it can be combined with ecological models or remote sensing & ML to project carbon sequestration.

Problem

- Machine learning has become an increasingly utilized method of classifying land cover.

- Although many methods use spatial data from remote sensing images, here only cartographic variables are used to predict tree species.

Data from: http://archive.ics.uci.edu/ml/datasets/Covertype

Data

Features Variables:
Elevation, Aspect, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_Roadways, Hillshade_9am, Hillshade_Noon, Hillshade_3pm, Horizontal_Distance_To_Fire_Points, Soil_Type (40 binary columns)

Target Variables:
Cover_Type (7 types of trees)

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

Expected Accuracy

“The overall mean absolute classification accuracy for the neural network method was 70.52%, with a 95% confidence interval of 70.26% to 70.80%.” -- Blackard, 1998

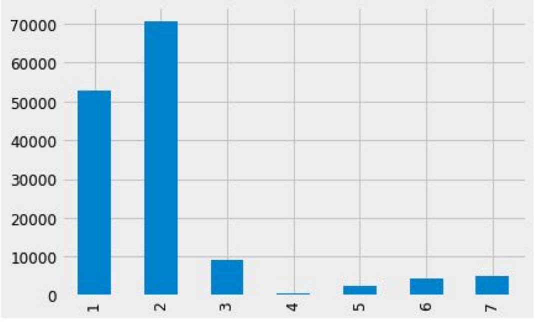
Blackard, Jock A. 1998. "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types." Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado. 165 pages.

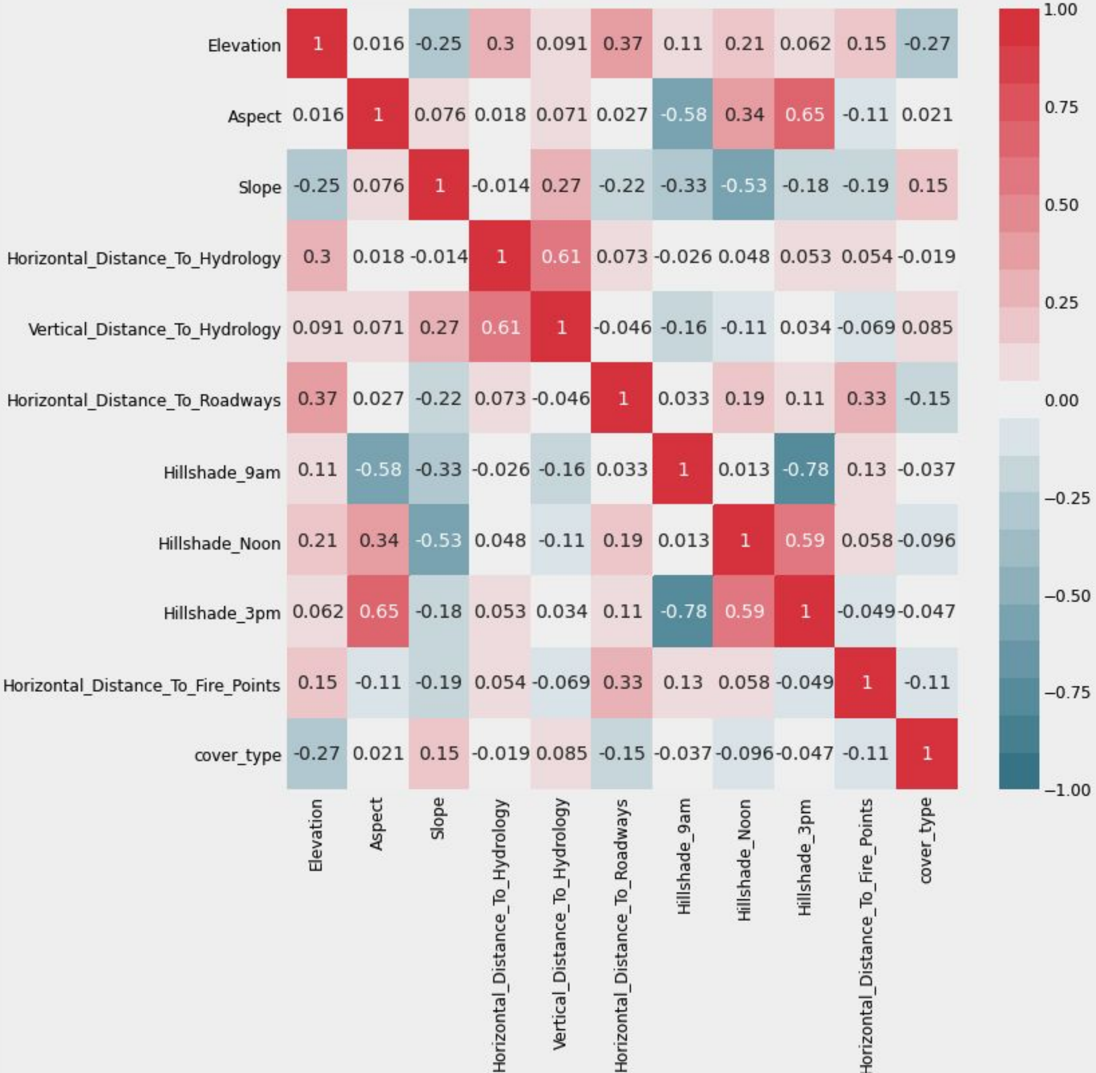
Limitations:

- This may be an oversimplification of the ecological processes underlying tree species distribution and growth
- This data is specific to Colorado & only 7 tree species. The models here wouldn’t be applicable to tropical forests

EDA

- 581,012 rows and 51 columns
    - Subsetted ¼ of the data
    - 145,253 rows and 51 columns
- Numeric features were not normally distributed
- Target feature had significantly more observations in class 1 & 2





	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade_Noon	Hillshade_3pm	Horizontal_Distance_To_Fire_Points	cover_type
Elevation	1	0.016	-0.25	0.3	0.091	0.37	0.11	0.21	0.062	0.15	-0.27
Aspect	0.016	1	0.076	0.018	0.071	0.027	-0.58	0.34	0.65	-0.11	0.021
Slope	-0.25	0.076	1	-0.014	0.27	-0.22	-0.33	-0.53	-0.18	-0.19	0.15
Horizontal_Distance_To_Hydrology	0.3	0.018	-0.014	1	0.61	0.073	-0.026	0.048	0.053	0.054	-0.019
Vertical_Distance_To_Hydrology	0.091	0.071	0.27	0.61	1	-0.046	-0.16	-0.11	0.034	-0.069	0.085
Horizontal_Distance_To_Roadways	0.37	0.027	-0.22	0.073	-0.046	1	0.033	0.19	0.11	0.33	-0.15
Hillshade_9am	0.11	-0.58	-0.33	-0.026	-0.16	0.033	1	0.013	-0.78	0.13	-0.037
Hillshade_Noon	0.21	0.34	-0.53	0.048	-0.11	0.19	0.013	1	0.59	0.058	-0.096
Hillshade_3pm	0.062	0.65	-0.18	0.053	0.034	0.11	-0.78	0.59	1	-0.049	-0.047
Horizontal_Distance_To_Fire_Points	0.15	-0.11	-0.19	0.054	-0.069	0.33	0.13	0.058	-0.049	1	-0.11
cover_type	-0.27	0.021	0.15	-0.019	0.085	-0.15	-0.037	-0.096	-0.047	-0.11	1

Results: Logistic Regression

Simple Logistic Regression

Accuracy:

Train : 0.70

Test: 0.70

** Null model would predict an accuracy of 0.49 based on the most frequent cover type

GridSearchCV and K-Fold Cross Validation

Accuracy:

Train : 0.70

Test: 0.70

*precision & recall scores were significantly lower for less frequent observations

L2 penalty (Lasso Regression)

Accuracy:

Train : 0.70

Test: 0.70

Feature importance →

Results: Random Forest Classifier

Random Forest

Hyper Parameter Tuning

Accuracy:

Train : 1.0

Test: 0.91

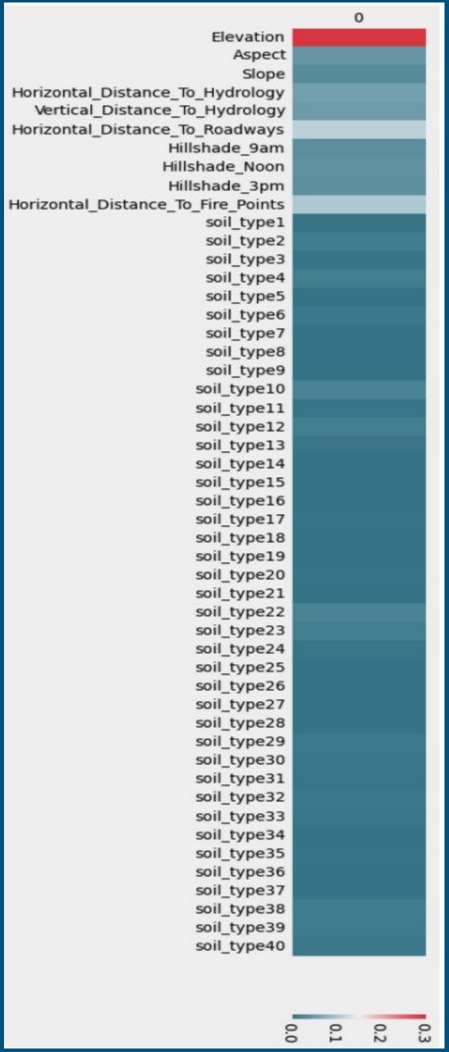
*Overfitting the model (high variance)

Accuracy:

Train : 0.99

Test: 0.92

*Overfitting the model (high variance)



Discussion

Tuning more or other models:

- Neural Networks

Recommendations:

- Using remote sensing data can likely predict cover type better than cartographic data
- It could be that a combination of remote sensing data & cartographic data fed into ML algorithms could produce accurate classification of tree cover type