Forest Cover Type Prediction

Christen Ye & Mariana Chen & Anastasia Ivanova

March 14, 2023

Agenda

- Introduction
- Source of Dataset
- Summary of relevant variables
- Data Exploratory Analysis
- Next Steps

Introduction

Nowadays, one of the most important global concern is forest deforestation and its monitoring. Understanding the different types of forest cover can aid in predicting forest ecosystem response to climate change and other disturbances. It is also useful in natural resource managements, such as for timber harvesting, wildlife habitat management and recreation planning.

In this project, we aim to apply Classification Tree, Random Forest and Naïve Baye's Classifier to predict the forest cover types.

Source of Dataset

The dataset called “Forest Cover Type Dataset” is from Kaggle.

- It contains 581,012 observations of cover types from four wilderness areas located in the Roosevelt National Forest of northern Colorado.
- Explanatory 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, Wilderness_Area #1 - Wilderness_Area #4, Soil_Type #1 - Soil_Type #40
- Response variable: Cover_Type (1 - Spruce/Fir, 2 - Lodgepole Pine, 3 - Ponderosa Pine, 4 - Cottonwood/Willow, 5 – Aspen, 6 - Douglas-fir, 7 - Krummholz)

Summary of Relevant Variables

	count	mean	std	min	25%	50%	75%	max
Elevation	581012.0	2959.365301	279.984734	1859.0	2809.0	2996.0	3163.0	3858.0
Aspect	581012.0	155.656807	111.913721	0.0	58.0	127.0	260.0	360.0
Slope	581012.0	14.103704	7.488242	0.0	9.0	13.0	18.0	66.0
Horizontal_Distance_To_Hydrology	581012.0	269.428217	212.549356	0.0	108.0	218.0	384.0	1397.0
Vertical_Distance_To_Hydrology	581012.0	46.418855	58.295232	-173.0	7.0	30.0	69.0	601.0
Horizontal_Distance_To_Roadways	581012.0	2350.146611	1559.254870	0.0	1106.0	1997.0	3328.0	7117.0
Hillshade_9am	581012.0	212.146049	26.769889	0.0	198.0	218.0	231.0	254.0
Hillshade_Noon	581012.0	223.318716	19.768697	0.0	213.0	226.0	237.0	254.0
Hillshade_3pm	581012.0	142.528263	38.274529	0.0	119.0	143.0	168.0	254.0
Horizontal_Distance_To_Fire_Points	581012.0	1980.291226	1324.195210	0.0	1024.0	1710.0	2550.0	7173.0
Wilderness_Area	581012.0	2.114462	1.061295	1.0	1.0	2.0	3.0	4.0
Soil_Type	581012.0	24.362443	9.485405	1.0	20.0	29.0	31.0	40.0
Cover_Type	581012.0	2.051471	1.396504	1.0	1.0	2.0	2.0	7.0

Figure 1. Summary Table for Explanatory Variables

Exploratory Data Analysis 1

Figure 2. Histogram of Numeric Variables

Exploratory Data Analysis 2

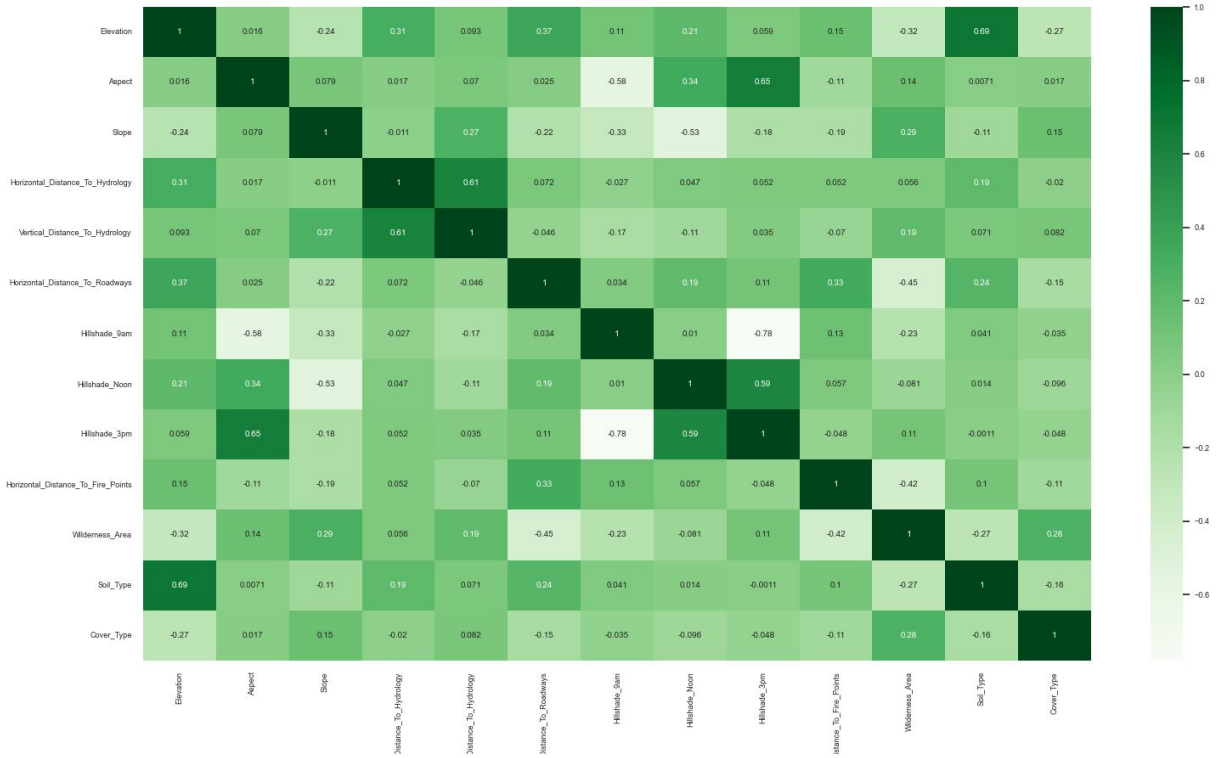


Figure 3. Correlation Plot between Variables

Exploratory Data Analysis 3

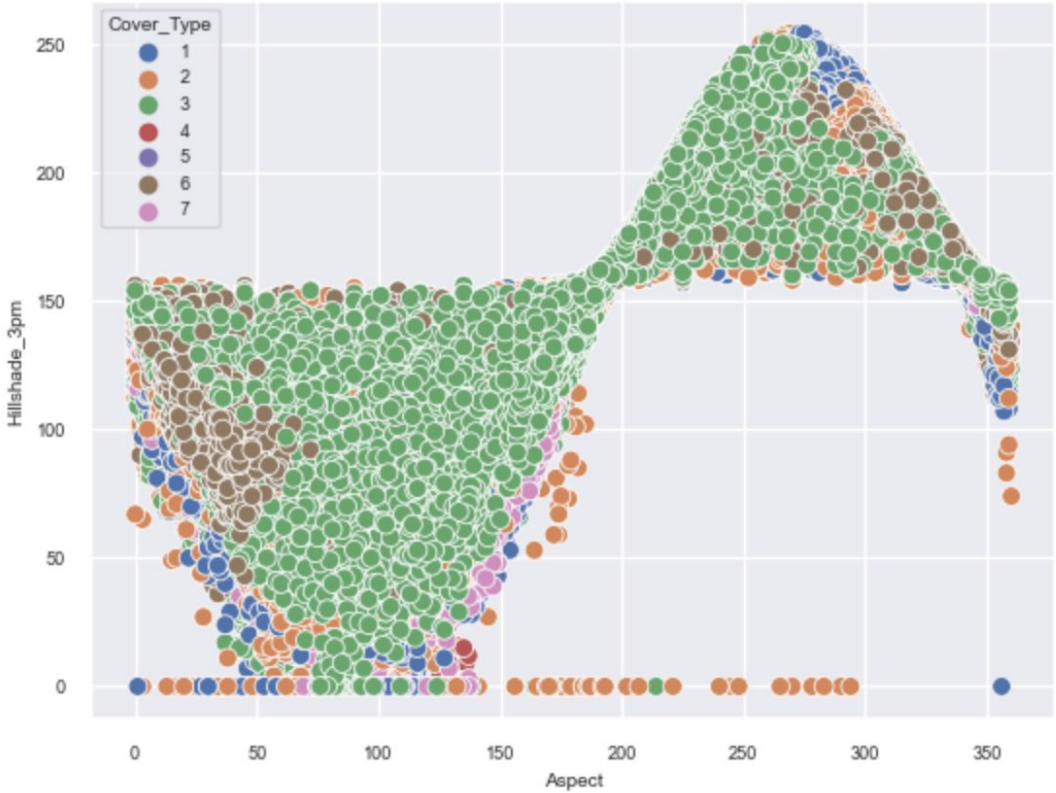


Figure 4. Scatter Plot between Aspect and Hillshade_3pm

Exploratory Data Analysis 4

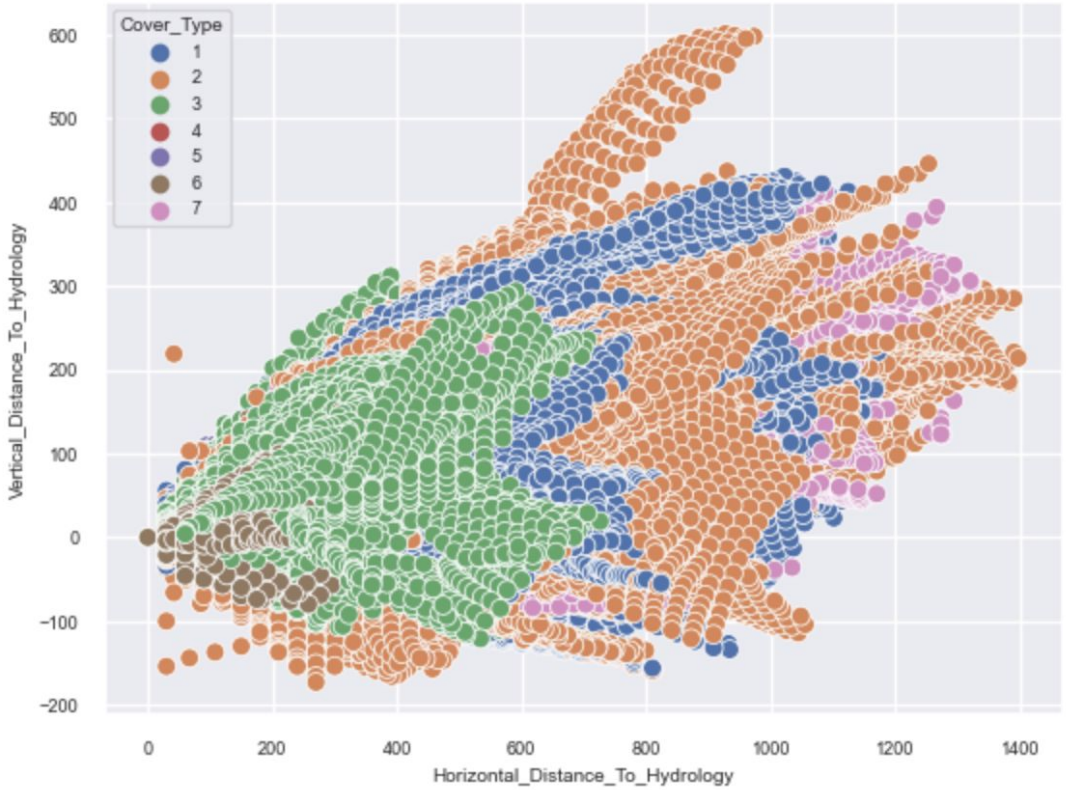
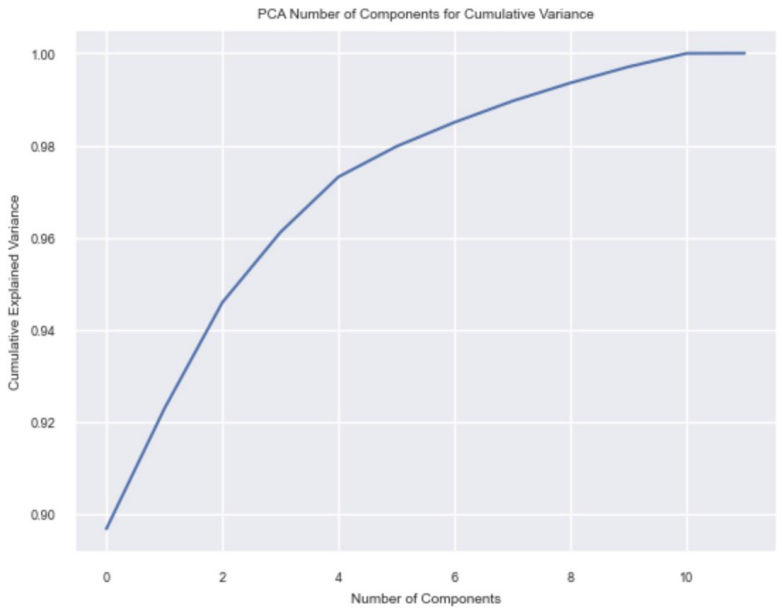
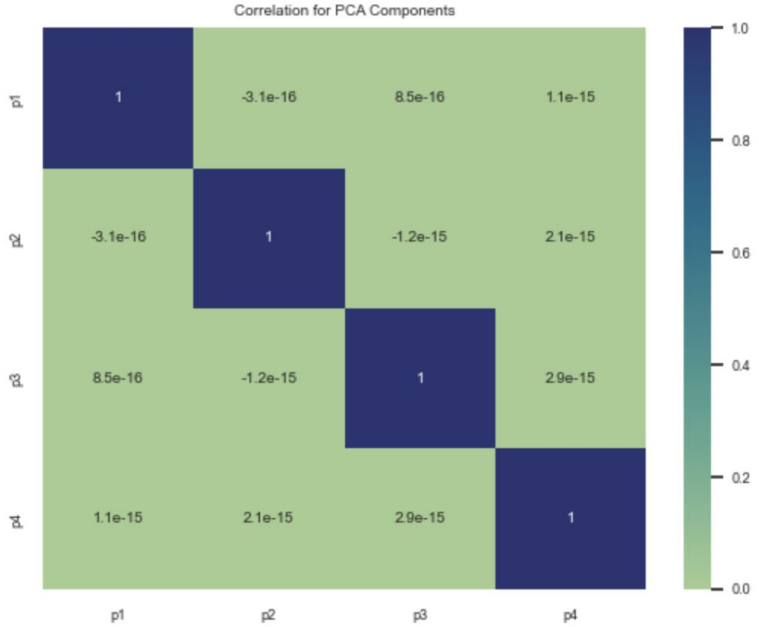


Figure 5. Scatter Plot between Vertical and Horizontal_Distance_To_Hydrology

Principal Component Analysis





Exploratory Data Analysis 5

Cover_Type	Counts	Proportions
1	211840	36.46%
2	283301	48.76%
3	35754	6.15%
4	2747	0.47%
5	9493	1.63%
6	17367	2.99%
7	20510	3.53%

Table1. Summary Table for Target Variable

Exploratory Data Analysis 6

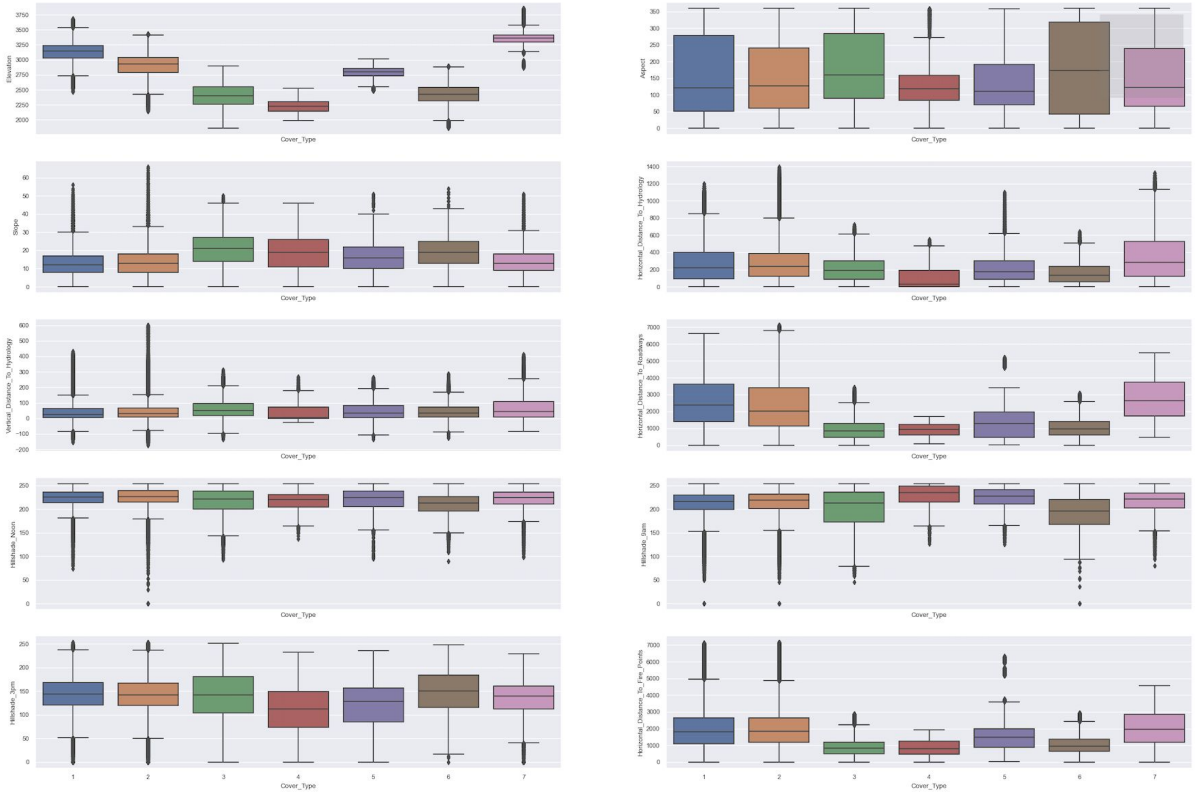


Figure 4. Side-by-side Boxplots for Cover Type by Features

Next Steps

- Feature selection: model-based selection, recursive feature elimination and forward selection.
- Comparison of different models (Classification Tree, Random Forest and Naïve Baye's Classifier) using k-fold cross validation.
- Performance measurements: out-of-sample accuracy using 50% and 80% prediction intervals, AUC and F1 score.