Neil Strugnell
4/14/2024
MSBA 8490

Week 3 – Initial Model Findings

Overview

For my second Case Study, I chose to look at the forest covertype dataset. The stakeholders for this project are the Colorado Department of Natural Resources requesting the analysis, and through the DNR residents of the state of Colorado and businesses who stand to benefit from having a better understanding of the condition of state forests. The DNR did not provide a specific timeline for their request, so the project could take anywhere from a day or two to a few weeks to complete depending on the amount of follow-up analysis the department requests after receiving initial results. The DNR did not provide any hypotheses or prior data of their expectations for changes in covertype over time.

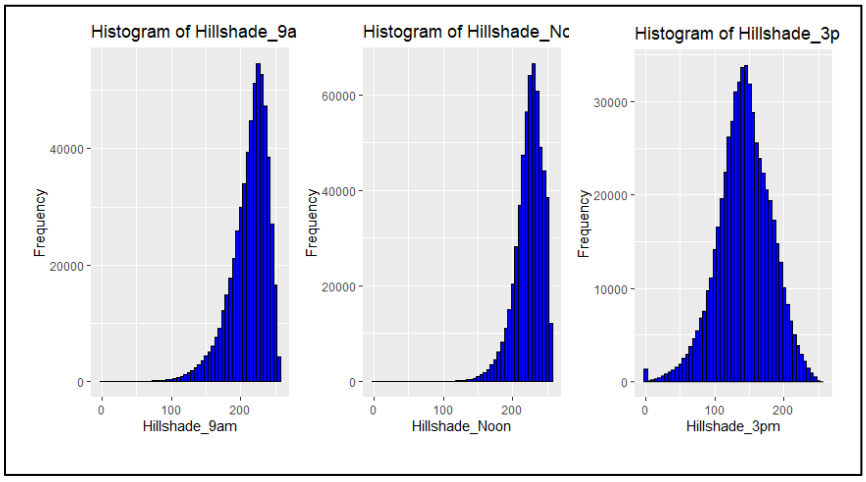
The Colorado DNR is requesting the project because they want to get a better sense of the factors that causes changes in covertype over time, so that they can predict future changes in the region's forests. Understanding the patterns of regional growth over time will help the department maintain healthy local ecologies, and reduce the risk catastrophes such as wildfire or extensive flooding.

Data and Model

The covertype dataset contained 581,012 observations of distinct 30 x 30 meter cells taken from four wilderness areas within Roosevelt National Forst in northern Colorado. Minimal background information was given on the four observed areas, but the dataset did include notes on the relative elevation and major trees species of each region. The dataset was presented by UC Irvine cleaned and ready for analysis with no missing values or duplicated observations, and all observations contained exclusively integer values that imported into R correctly as numeric datatype. The variable of interest was Cover_Type, an integer value of 1 to 7 indicating the primary type of tree growth in each observed cell from spruce or fir, lodgepole pine, ponderosa pina, cottonwood/willow, Douglas-fir, or Krummholz.

After verifying the integrity of the data, I created several graphs to perform a visual inspection of important aspects of the data. Histograms of the hillshade variables for 9:00 AM, noon, and 3:00 PM showed that most of the surveyed regions are in shade in the morning and around noon, with hillshade cover lowest in the afternoon.

Histograms of Hillshade



I investigated summary statistics for all variables related to elevation, as well as the hillshade in each observed area through the day and the distance between each cell and hydrology, roadways, and fire points. There is a range of nearly 2,000 meters between the lowest and highest observed elevations, and negative observed values for the vertical distance to hydrology variable for observed regions with elevation below a certain value, though the notes accompanying the dataset did not indicate what was used as a 0 value for this variable.

Summary Stats

	Feature	Count	Mean	SD	Min	X1st.Qu.	Median	X3rd.Qu.	Max
Elevation	Elevation	581012	2959.37	279.98	1859	2809	2996	3163	3858
Aspect	Aspect	581012	155.66	111.91	0	58	127	260	360
Slope	Slope	581012	14.10	7.49	0	9	13	18	66
Horizontal_Distance_To_Hydrology	Horizontal_Distance_To_Hydrology	581012	269.43	212.55	0	108	218	384	1397
Vertical_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	581012	46.42	58.30	-173	7	30	69	601
Horizontal_Distance_To_Roadways	Horizontal_Distance_To_Roadways	581012	2350.15	1559.25	0	1106	1997	3328	7117
Hillshade_9am	Hillshade_9am	581012	212.15	26.77	0	198	218	231	254
Hillshade_Noon	Hillshade_Noon	581012	223.32	19.77	0	213	226	237	254
Hillshade_3pm	Hillshade_3pm	581012	142.53	38.27	0	119	143	168	254
Horizontal_Distance_To_Fire_Points	Horizontal_Distance_To_Fire_Points	581012	1980.29	1324.20	0	1024	1710	2550	7173

The notes on the dataset contained little information on the 40 soil type variables included, so I looked at a correlation table of all soil types and the Cover_Type variable to check for tentative relationships between each soil type and the Cover_Type. I observed significant positive correlations with soil types 9, 37, 38, and 39, and significant negative correlations with soil types 21, 22, and 28, and included all seven of these variables in my model along with the continuous variables included in my summary stats.

Using the above 10 variables plus the 7 selected soil type variables, I ran a logistic regression to investigate the impact of each variable on Cover_Type.

Results

The model results from my initial analysis were inconclusive as almost all variables showed significant relationships with the Cover_Type variable of interest. Positive and negative coefficients on these variables show trends between each variable and the resulting covertype, so increases in elevation, morning and late afternoon hillshade, and all the soil types negatively correlated with Cover_Type indicate that these variables are more likely to produce regions with spruce/fir or pine tree covertypes, while positive coefficients are more likely to produce cottonwood or willow trees, or the Douglas-fir or Krummholz covertypes coded at the top end of the variable's range.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.531e+01	2.357e-01	64.971	< 2e-16	***
Elevation	-6.746e-03	2.191e-05	-307.873	< 2e-16	***
Aspect	5.219e-04	3.973e-05	13.138	< 2e-16	***
Slope	2.115e-02	1.199e-03	17.642	< 2e-16	***
Horizontal_Distance_To_Hydrology	1.681e-03	2.034e-05	82.619	< 2e-16	***
Vertical_Distance_To_Hydrology	2.499e-03	7.767e-05	32.174	< 2e-16	***
Horizontal_Distance_To_Roadways	3.174e-05	2.228e-06	14.245	< 2e-16	***
Hillshade_9am	-4.513e-03	1.384e-03	-3.261	0.00111	**
Hillshade_Noon	3.256e-02	1.132e-03	28.770	< 2e-16	***
Hillshade_3pm	-1.230e-02	1.132e-03	-10.859	< 2e-16	***
Horizontal_Distance_To_Fire_Points	7.111e-06	2.692e-06	2.642	0.00825	**
Soil_Type9	-1.334e+00	8.644e-02	-15.430	< 2e-16	***
Soil_Type21	-3.054e+00	1.772e-01	-17.241	< 2e-16	***
Soil_Type22	-9.842e-01	1.446e-02	-68.063	< 2e-16	***
Soil_Type28	7.791e-01	1.591e-01	4.898	9.68e-07	***
Soil_Type37	1.500e+01	1.829e+01	0.820	0.41218	
Soil_Type38	1.132e+00	1.821e-02	62.186	< 2e-16	***
Soil_Type39	1.038e+00	1.973e-02	52.617	< 2e-16	***

Conclusion

The low amount of information provided about the 40 soil type variables included in the dataset, as well as the absence of any categorical variables to use as reference points, made this dataset somewhat difficult to analyze. The biggest missing piece is the lack of longitudinal data, which would provide vital details on changes in covertype over time, ideally or an annual or even decadal basis, though there are admittedly difficulties with gathering observations for such a wide area over an extended period.

Lacking the capacity for long term observations, I would recommend the Colorado DNR supplement the provided data with observations from ecologically similar regions from other parts of the state or comparable parks from nearby states in the Rocky Mountain region.