Shiping Li

FOREST COVER TYPE CLASSICFICATION

1. INTRODUCTION

a) Background

Forests cover about 30% of our earth's land surface and are crucial to our fundamental life. It purifies the air we breathe every day, supplies tremendous natural resources to all of us, mitigates climate change, and reduces the risk of soil erosion and other natural disasters. All these services provided by the forest ecosystem enable our daily survival and sustainable development in all aspects. Therefore, protecting and learning the forest ecosystem will always be an endless subject, especially it is being destroyed unprecedentedly by human activities. Identifying the non-urban forest cover type can tell us how the surrounding environment interacts with a particular forest cover type and what elements decide the predominant kind of cover type on a site. It will provide the scientific basis for the governments and environmental organizations to monitor or reconstructing the pre-settlement of vegetation patterns.

b) Objective

The objective of the project is to use cartographic variables (as opposed to remotely sensed data) to classify 7 forest cover types. The study area includes 4 wilderness areas located in Roosevelt National Forest of northern Colorado, representing forests with minimal human-caused disturbances.

2. DATASET INFORMATION

The dataset was retrieved from UCI Machine Learning Repository containing 501812 observations. The observations are distributed in 4 different areas in Roosevelt National Forest of northern Colorado including:

- Rawah Wilderness Area (Area 1)
- Neota Wilderness Area (Area 2)
- Comanche Peak Wilderness Area (Area 3)
- Cache la Poudre Wilderness Area (Area 4)

Neota Wilderness Area has the mean highest elevational value, followed by Rawah and Comanche Peak Wilderness Areas. Cache la Poudre Wilderness Area has the lowest mean

elevational value among the four areas. Each observation is a 30m x 30m patch of forest that is classified as one of the seven cover types, determined by US Forest Service (USFS) Region 2 Resource Information System (RIS) data. The seven cover types are represented by numbers from 1 to 7 in our dataset: (1) Spruce/Fir (2) Lodgepole Pine (3) Ponderosa Pine (4) Cottonwood/Willow (5)Aspen (6) Douglas-fir (7) Krummholz

Figure 1: Seven Forest Cover Types From (1) Spruce/Fir to (7)Krummholz

The dataset consists of 54 variable. To better understand, we can divide the variables into 5 types (in addition to our target variable Cover_Type). The first type is in terms of **terrain** features of an observed patch, including Elevation, Aspect, and Slope. The second type is in terms of **distance**, including Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_Roadways, and Horizontal_Distance_To_Fire. The third type is related to the **insolation duration**, including Hillshade_9am, Hillshade_Noon, Hillshade_3pm. The fourth type is regarding **the location** of an observed patch, including four different areas named from Wilderness_Area1 to Wilderness_Area4, and the fifth type is related to 40 **soil types** named from Soil_Type1 to Soil_Type40. The first three types are continuous variables, while the last two types **location** and **soil types** are binary variables with values 0 and 1, where 0 = 'absence' and 1 = 'presence' for the corresponding attribute in a column. Our target value Cover_Type is a categorical variable with integer numbers ranging from 1 to 7, where each number represents a unique forest cover type.

3. DATA WRANGLING

The raw dataset obtained from UCI Machine Learning Repository was very clean and tidy, so it remained intact at the end of this step. The dataset consists of 581012 rows and 54 columns. No missing values were found.

In the summary table, I found most features has positive values, except Vertical_Distance_To_Hydrology. The table showed the min value of Vertical_Distance_To_Hydrology was -173, which looked suspicious. According to the description provided by the data contributor, Vertical_Distance_To_Hydrology explaines the

vertical distance from the sample forest to its nearest surface water, which might explain why its value could be under 0 when considering the direction. The elevation of the patch of forest could be either higher or lower than the elevation of the surface water, resulting in positive or negative values. In addition, near 20% values were negative, ranging from -173 to 0. This information should give strong evidence that the negative data were not caused by input error.

4. DATA EXPLORATORY ANALYSIS

a) Distribution of the Target Value

The following figure showes that the seven forest cover types are distributed extremely uneven. Cover_Type1 and Cover_Type2 both have counts more than 200000, accounting for majority of the dataset. Cover type4 has the least number of counts, far below than 10000.

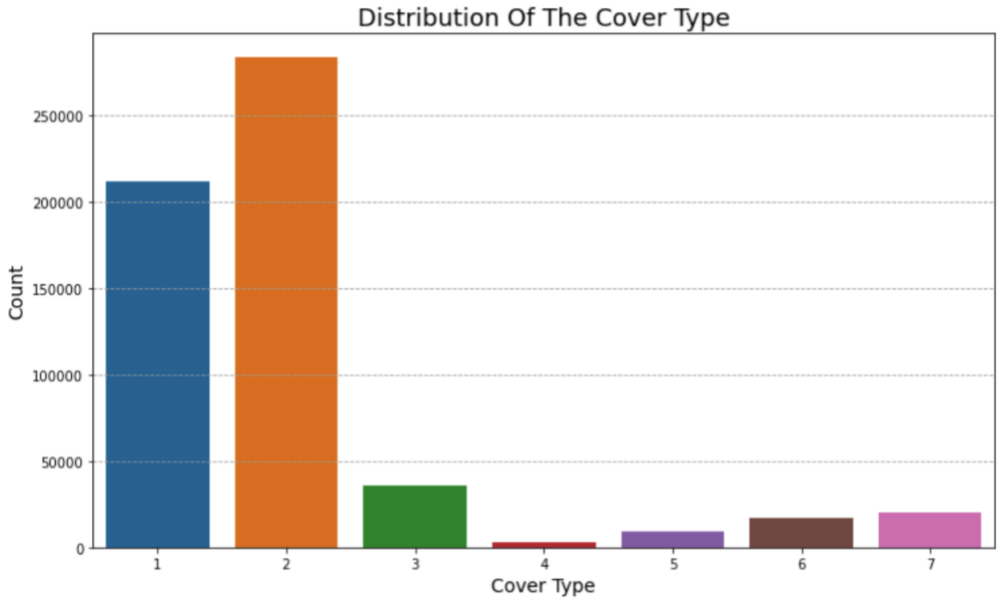


Figure 2: Counts for Each Cover Types

Here is the proportion of each cover types in our data set. Cover type 1 and cover type 2 account for the more than 85% of the total observations while cover type 4 only accounts for less than 0.47%.

- Cover Type 2 (Lodgepole Pine): 48.76%
- Cover Type 1 (Spruce/Fir): 36.46%
- Cover Type 3 (Ponderosa Pine): 6.15%
- Cover Type 7 (Krummholz): 3.53%
- Cover Type 6 (Douglas-fir): 2.99%
- Cover Type 5 (Aspen): 1.63%
- Cover Type 4 (Cottonwood/Willow): 0.47%

The problem of umbalance classes will increase the difficulty in the identifying the minority from the majority. Such that, I addressed this issue by resampling the data, which will be discussed in the last section.

b) Main Forest Cover Type in Each Area

To explore the relationship between the four wilderness areas and cover types, I grouped the data into four by the factor of the wilderness area and plot the distribution of cover type. There seems to be some class separations across four wilderness areas. We can see that the first two main cover types are Cover_Type1 and Cover_Type2 in Wilderness_Area 1, 2, and 3 respectively. The first two main cover types in Wilderness_Area4 are cover type 3 and cover type 6. Only cover type 2 is presented in all four wilderness areas. Then cover type 2 is very likely to be more adaptive to varying environments.

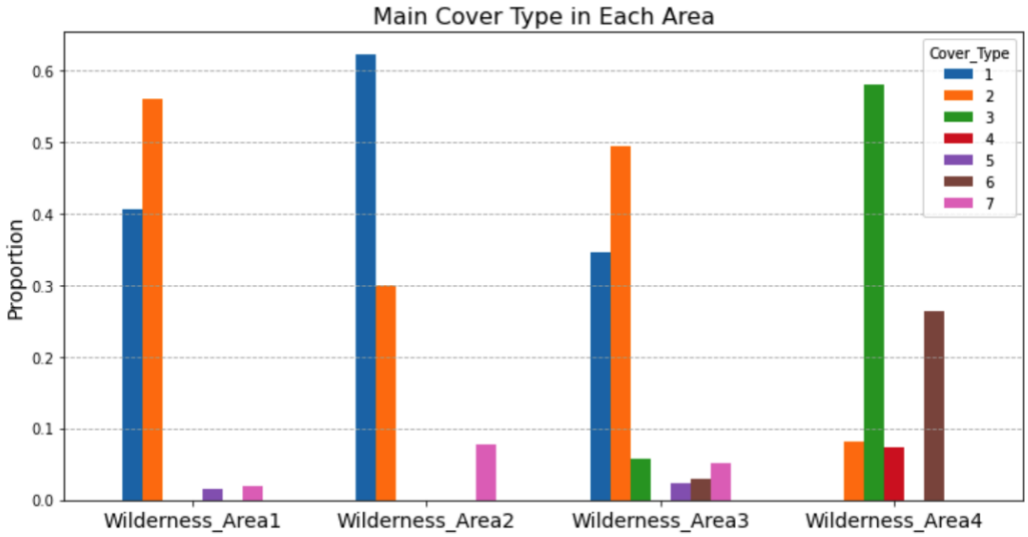
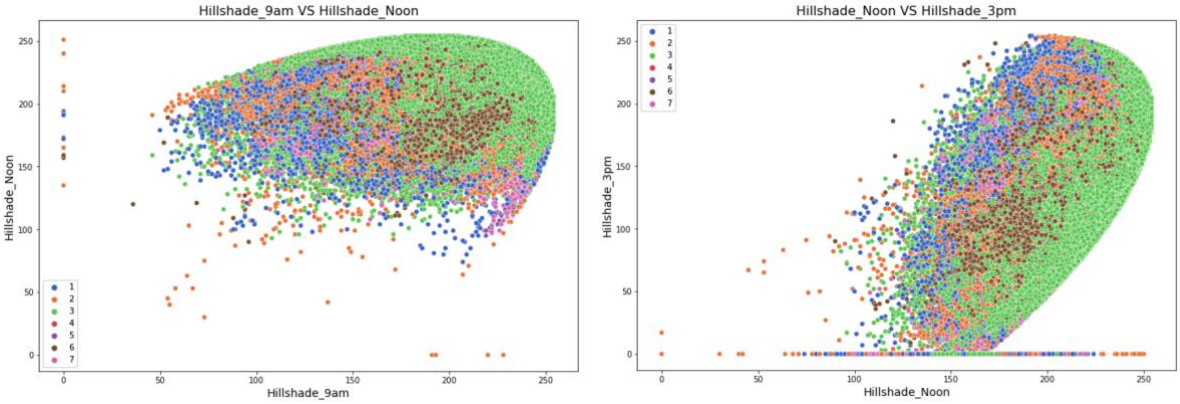


Figure 3: Distribution of the Cover Type Across 4 Wilderness Areas

To find out the potential factors affecting the main forest cover types, we could find out how the environmental conditions of these four areas distinct from others. In particular, the Wilderness_Area4 is the only one having different main cover types, so we should expect that Wilderness_Area4 has some strong characteristics that the other do not have. One thing we've mentioned above was the elevation of Wilderness_Area4 has the lowest elevation compared to the other areas. Its mean elevation is close 2300m, far less than Wilderness_Area1 and Wilderness_Area3's 3000m and Wilderness_Area2's 3200m. Cover_Type3 and 6 might probably favor middle elevation over high elevation. Another thing we found was that Wilderness_Area4 has the shortest average horizontal distance to fire points, given its mean equals to 779m, at least twice smaller than the others. This may suggest that cover type 3 and 5 grow in a drier environment and they are relatively drought tolerant.

c) Insolation Duration and Cover Types

In figure 4, the scatters plot shows the hillshade index of the observations at various times.The hillshade index ranges from 0 to 250, the higher value the more sunlight. All these seven cover types seem to favor the sunlight over the shade. Cover_Type3 and Cover_Type2 spread in a wider range, which may suggest they favor the sunlight but are shade tolerant meanwhile.



(a) (b)

Figure 4: Hillshade Index at Various Time

d) Elevation and Cover Types

The distribution of elevation across seven cover types is significantly distinct from others, with varying range and variance! Cover_type7 grows at the highest elevation overall, followed by Cover_type1 and Cover_type2. Cover_type4 grows at the lowest elevation. Cover types growing at higher elevations may imply the characteristic of cold tolerance, vice versa. In addition, Cover_Type4, Cover_Type5 and Cover_Type7 spread in narrower ranges, indicating that they are less adaptable to environmental changes. In short, elevation should be a very important factor to classify the cover types.

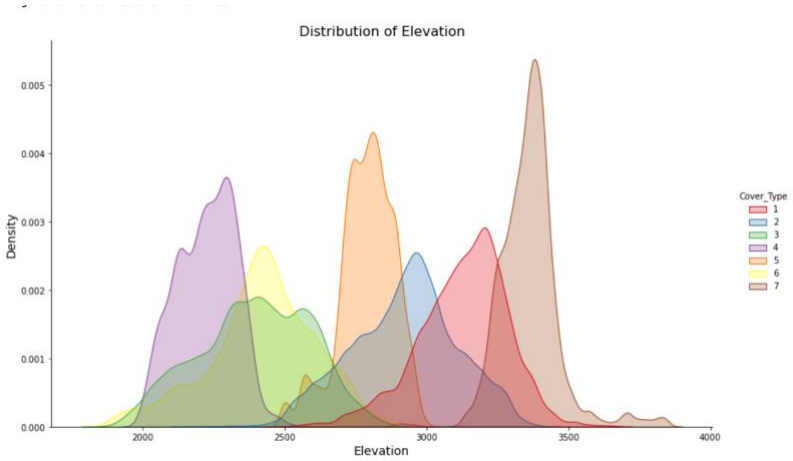


Figure 5: Distribution of Elevation Across 7 Cover Types

e) Distance To Hydrology and Cover Type

Based on the given horizontal and vertical distance to hydrology, I created a new feature of Euclidean distance. The following figure is the distribution of Euclidean distance to Hydrology across different cover types.

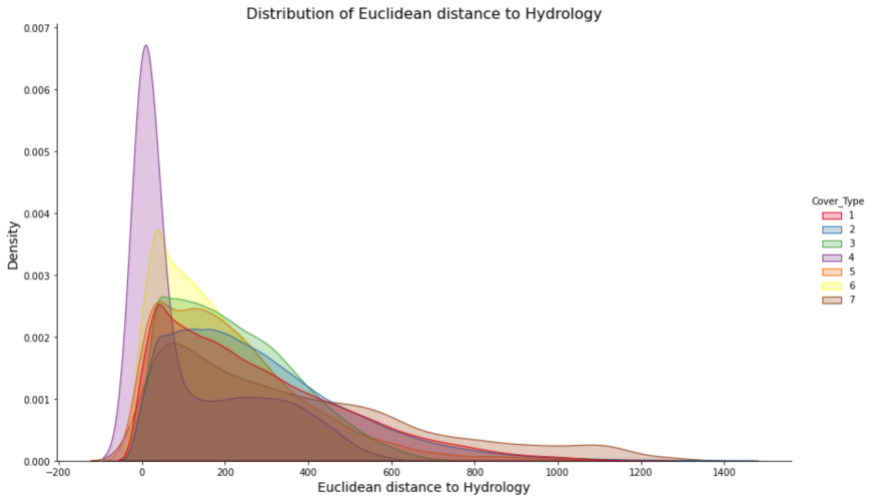


Figure 6: Distribution of Euclidean Distance to Hydrology Across 7 Cover Types

All the distributions are right-skewed, but strong varieties are still present among them. Cover_Type4 shows a higher need for water, and its peak is significantly higher than others. Cover_Type7 may be more drought enduring due to its wider spread. Overall, all cover types are water-loving because their peaks land between 0 - 200 meters from the water source.

5. MODEL SELECTION

a) Preparations for Modeling

Before training the data, I split 64% of data into the training set, 16% into the validation set, and 29% into the test set. I also performed target encoding technique on our categorial features Wilderness_Area and Soil_Type to reduce more than a half number of columns, which helped reduce the processing time of our baseline logistic regression model. The metric I choose for model evaluation is macro F1-score. The reason why macro F1-score is good option for case is it gives the same importance to each class. It will give a low score for models that only perform well on majority classes while perform poor on the minority classes. Meanwhile, it guarantees precision and recall balance.

b) Candidate Models And Final Selection

We have 4 candidate models at the end. Two decision tree models, one without resampling and one with resampling, the same is for random forest models. The oversampling method we used was Borderline-SMOTE.

The classification reports shows that the two random forest models outperformed the decision trees models in accuracy, macro average f1-score, precision, and recall on the validation set (figure 7). Both of the random forest models have the same macro f1-score 0.91 and accuracy 0.93 on the validation set, comparing to 0.88 and 0.93 of the decision tree model without resampling and 0.88 and 0.91 of the decision tree model with resampling. Moreover, the random forest models seem to have a slightly better generalization ability on the unseen data since they have a smaller difference in macro f1 scores on the training set and the validation set. Lastly, I would choose the random forest model with resampling as our final model. Our goal is to correctly identify each cover type as much as possible regardless of its sample size, so we want to make sure the minority Cover_Type4, 5 and 6 can be successfully caught out from the majority cover types. The random forest model with resampling is more powerful in identifying the minority classes because the recalls scores of Cover_Type4 and Cover_Type5 are 0.05 and 0.09 higher than those of the other random forest model.

Without Resampling

Cover Type	Precision	recall	f1-score	support
1	0.91	0.89	0.90	33894
2	0.91	0.93	0.92	45328
3	0.92	0.93	0.92	5721
4	0.85	0.82	0.83	439
5	0.81	0.79	0.80	1519
6	0.88	0.86	0.87	2779
7	0.95	0.93	0.94	3282
Accuracy			0.93	92962
Macro avg	0.89	0.88	0.88	92962
Weighted avg	0.91	0.91	0.91	92962

With Resampling

Cover Type	Precision	recall	f1-score	support
1	0.91	0.89	0.90	33894
2	0.92	0.93	0.92	45328
3	0.92	0.93	0.92	5721
4	0.83	0.82	0.82	439
5	0.78	0.84	0.81	1519
6	0.86	0.86	0.86	2779
7	0.95	0.93	0.94	3282
Accuracy			0.91	92962
Macro avg	0.98	0.89	0.88	92962
Weighted avg	0.91	0.91	0.91	92962

(a) Model1: Decision Tree (w/o Resampling)

Cover Type	Precision	recall	f1-score	support
1	0.94	0.90	0.92	33894
2	0.92	0.96	0.94	45328
3	0.93	0.96	0.94	5721
4	0.90	0.84	0.87	439
5	0.91	0.73	0.81	1519
6	0.92	0.89	0.90	2779
7	0.97	0.93	0.95	3282
Accuracy			0.93	92962
Macro avg	0.93	0.89	0.91	92962
Weighted avg	0.93	0.93	0.93	92962

(b) Model2: Decision Tree (w/ Resampling)

Cover Type	Precision	recall	f1-score	support
1	0.95	0.90	0.92	33894
2	0.92	0.96	0.94	45328
3	0.94	0.95	0.94	5721
4	0.87	0.89	0.88	439
5	0.87	0.82	0.84	1519
6	0.92	0.89	0.90	2779
7	0.97	0.93	0.95	3282
Accuracy			0.93	92962
Macro avg	0.92	0.90	0.91	92962
Weighted avg	0.93	0.93	0.93	92962

(c) Model3: Random Forest (w/o No Resampling)

(d) Model4: Random Forest (w/ Resampling)

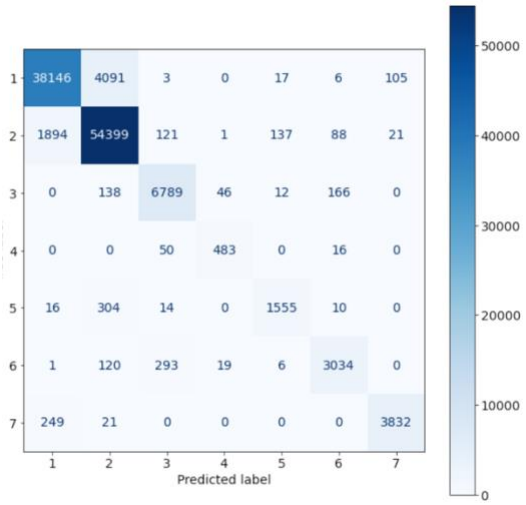
Figure 7: Results of Four Candidate Models

c) Results

At the end, we used our final model, the random forest model with resampling, to predict the test set and get the following result (Figure 8).

Cover Type	Precision	recall	f1-score	support
1	0.95	0.90	0.92	42368
2	0.92	0.96	0.94	56661
3	0.93	0.95	0.94	7151
4	0.88	0.88	0.88	549
5	0.90	0.82	0.86	1899
6	0.91	0.87	0.89	3473
7	0.97	0.93	0.95	4102
Accuracy			0.93	116203
Macro avg	0.92	0.90	0.91	116203
Weighted avg	0.93	0.93	0.93	116203

(a) Test Scores



(b) Confusion Matrix

Figure 8: Results of Test Set

The test scores are very satisfying and are close to the scores of the validation set. The macro average f1-score and accuracy on the test set are 0.91 and 0.93, which are same as the validation set. The performance on the minority Cover_Type4, 5 and 6 are very stable as well. The recall scores of Cover_Type4 and Cover_Type6 drop 0.01 and 0.02 respectively while the recall score on the Cover_Type5 remains unchanged. The performance on the majority classes are just as good as usual, with f1-scores, recalls, precisions are all above 0.90.