FOREST COVER CLASSIFICATION

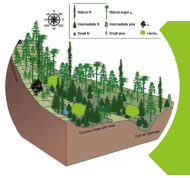
By Shiping Li

Why classify forest cover types?

Provide scientific basis for the future forest study.

Forest provides many foods and services.

Develop strategies for forest maintenance

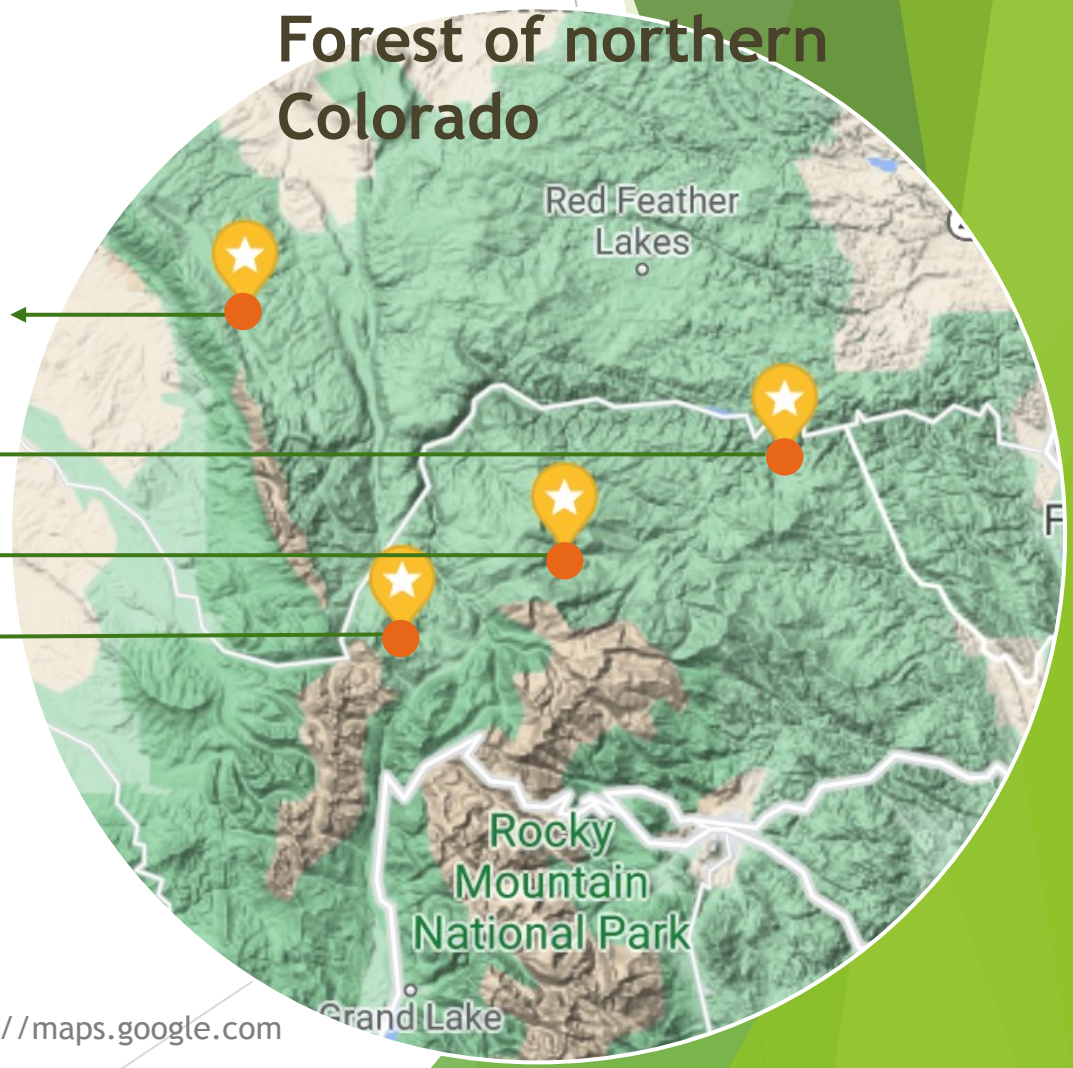


Reconstruct pre-settlement forest structure and distribution

Objective

The objective of the project is to classify 7 forest cover types.

Roosevelt National Forest of northern Colorado



Rawah Wilderness

Cache la Poudre Wilderness

Comanche Peak Wilderness

Neota Wilderness

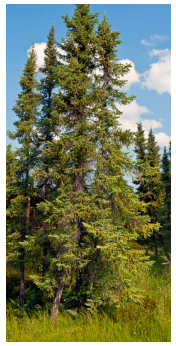
Map source: https://maps.google.com

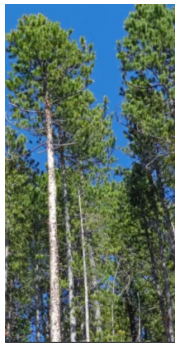
Data Info

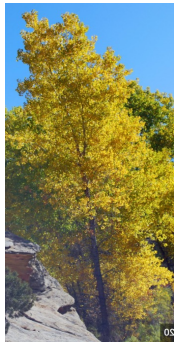
```
covtype.head()
```

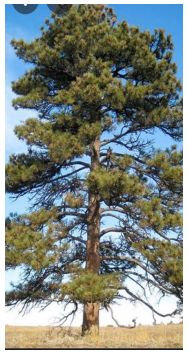
	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade_Noon
0	2596	51	3	258	0	510	221	232
1	2590	56	2	212	-6	390	220	235
2	2804	139	9	268	65	3180	234	238
3	2785	155	18	242	118	3090	238	238
4	2595	45	2	153	-1	391	220	234

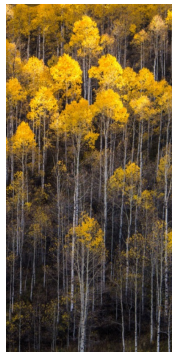
Each observation is a 30m x 30m patch of forest.
Target value: Cover Type

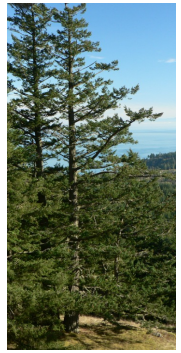

(1) Spruce/Fir


(2) Lodgepole Pine


(3) Ponderosa Pine


(4) Cottonwood/Willow


(5)Aspen


(6) Douglas-fir


(7) Krummholz

Data Info

Explanatory Variables

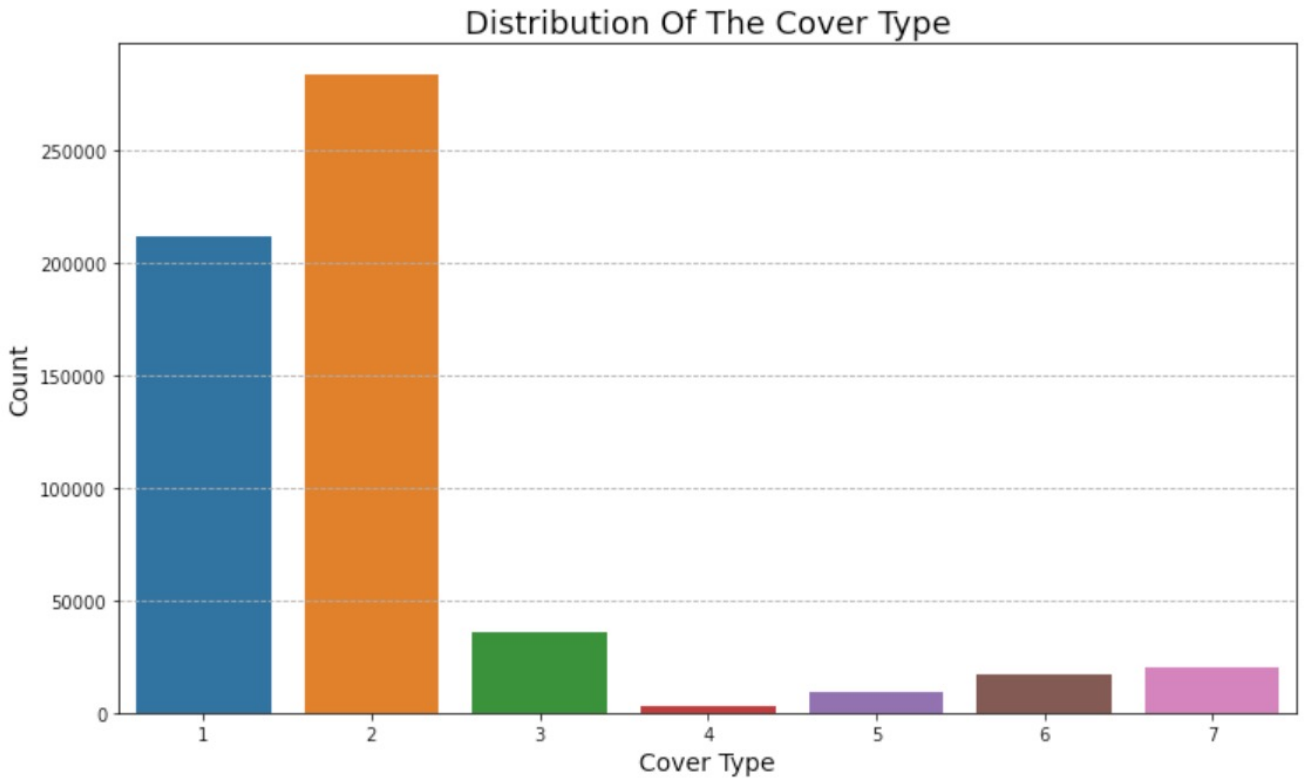
Terrain	Distance	Insolation Duration	Location	Soil Type
continuous	continuous	continuous	binary (categorical)	binary (categorical)
☐ Elevation (m)	☐ Horizontal Distance to Hydrology (m)	☐ Hillshade_9am (0 to 255 index)	☐ 4 Wilderness Areas (0 absence or 1 presence)	☐ 40 Soil Types (0 absence or 1 presence)
☐ Aspect (azimuth)	☐ Vertical Distance to Hydrology (m)	☐ Hillshade_Noon (0 to 255 index)		
☐ Slope (degree)	☐ Horizontal Distance to Roadways (m)	☐ Hillshade_3pm (0 to 255 index)		

Data Wranggling

- Download data from UCI Machine Learning Repository
- Original data had 501,812 rows and 54 columns
- No missing values were found
- Data was well structured and tidy.
- Found suspicious values: the Vertical_Distance_To_Hydrology had near 20% percent values negative; assume it took the direction into account, so keep them intact.

Data Exploratory Analysis

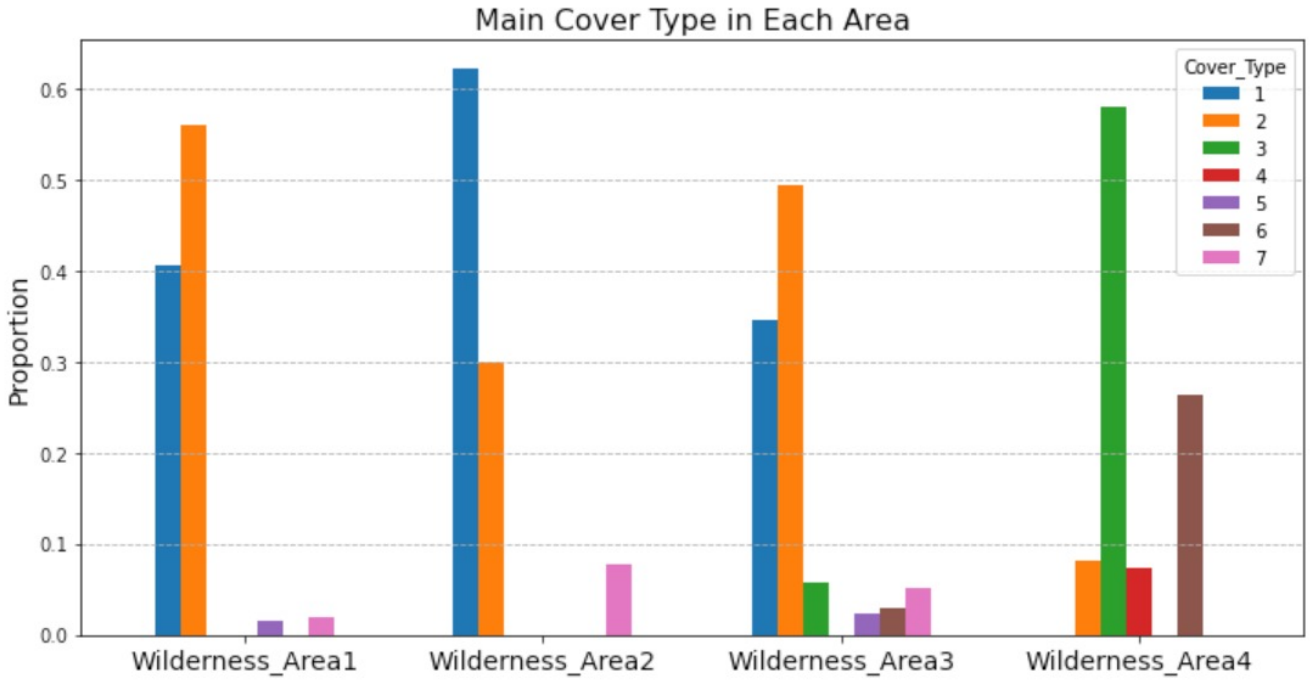
- Problem of Class Imbalance



- 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%

Data Exploratory Analysis

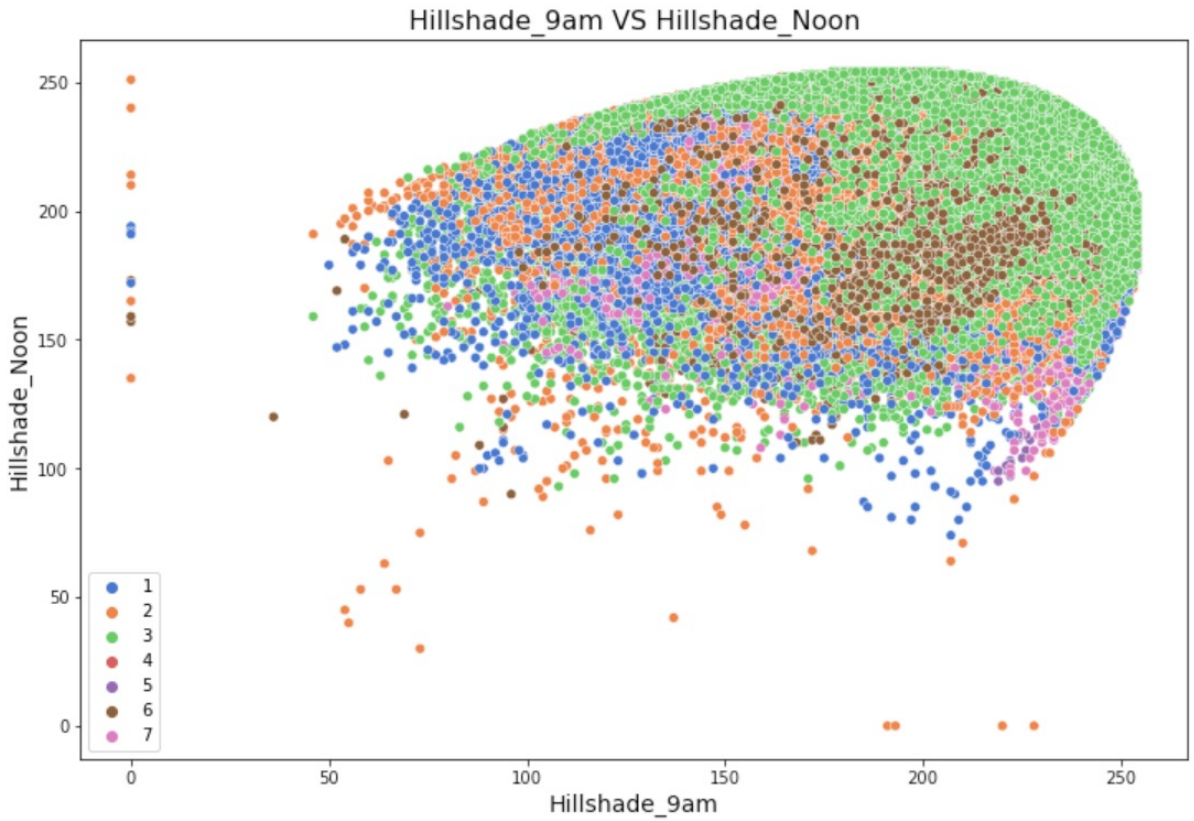
- Predominant Cover Type in Different Areas

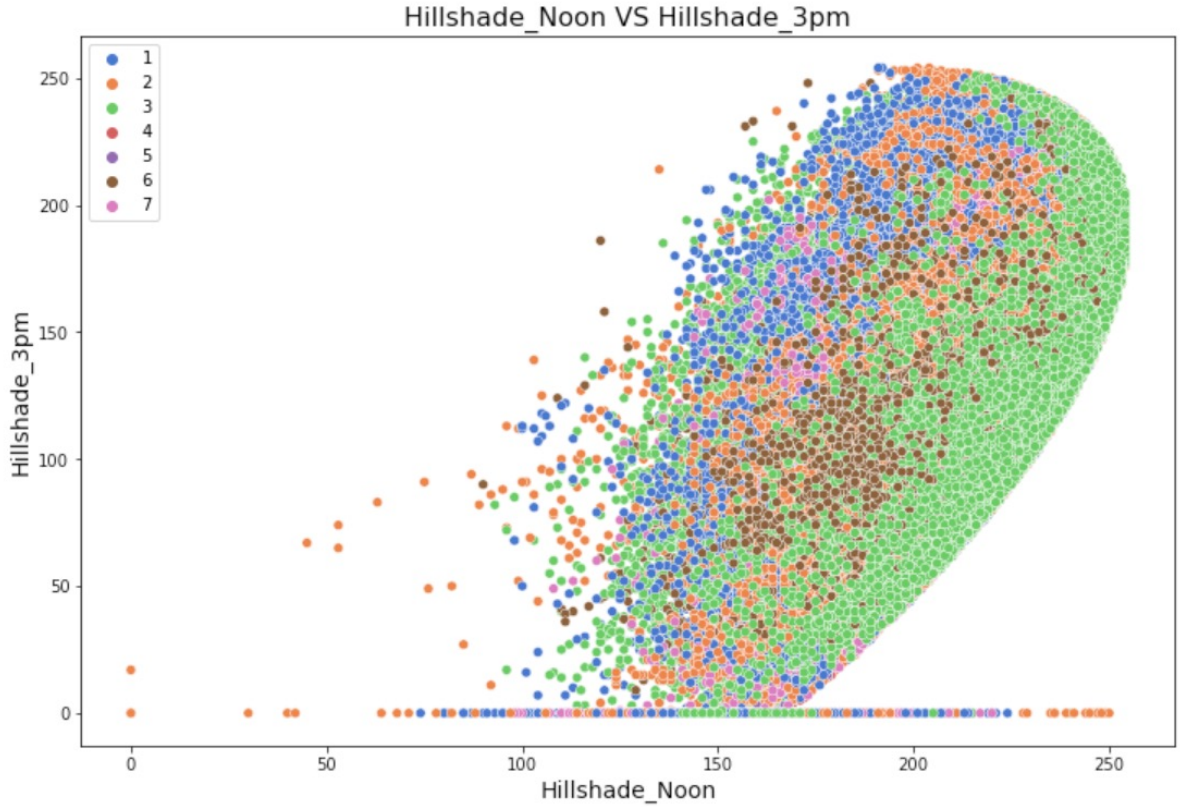


- In Wilderness Area 1, 2 , 3: Main cover types are 1 and 2
- In Wilderness Area 4: Main cover types are 3 and 6
- Only cover type 2 appears in all Wilderness Areas.(More adaptive?)

Data Exploratory Analysis

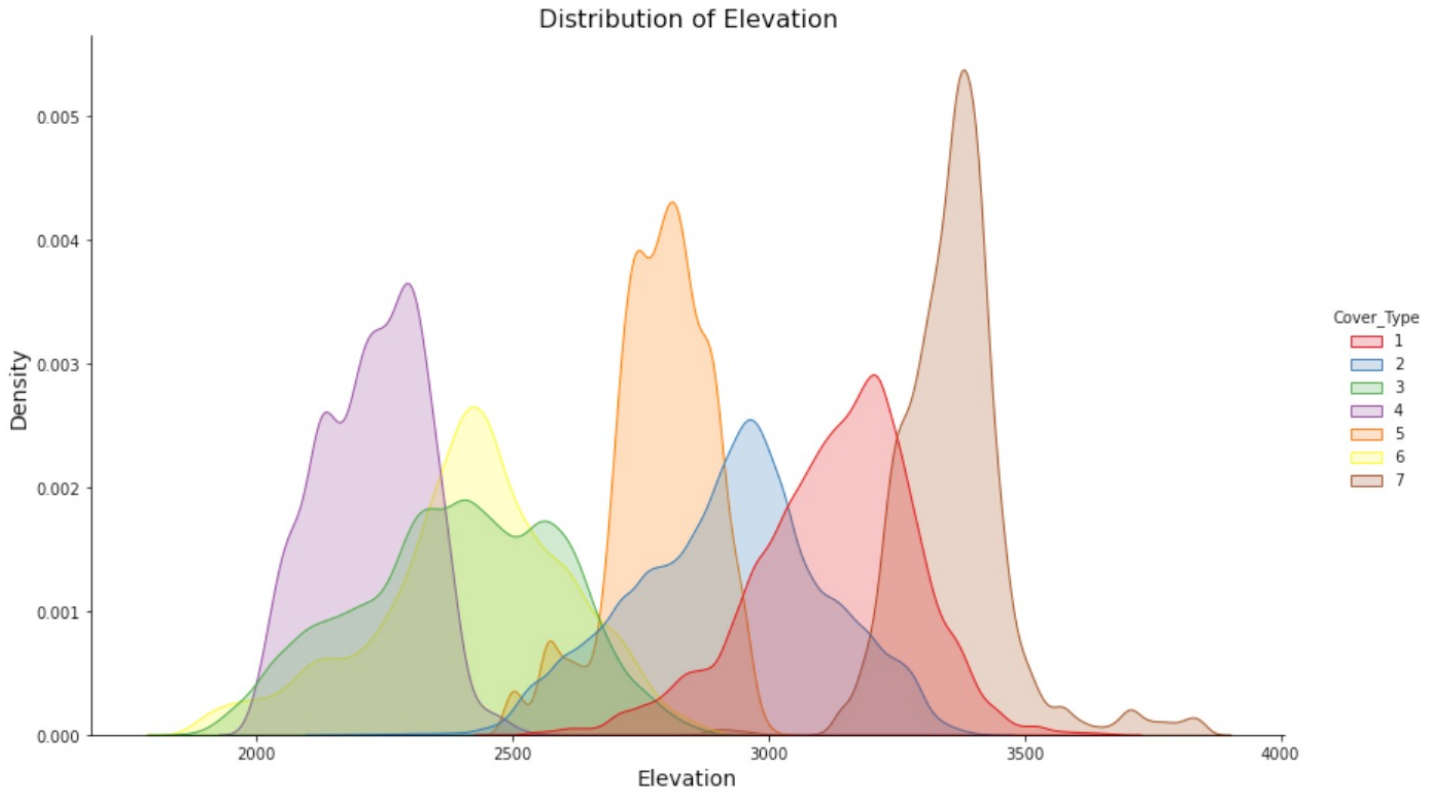
- Insolation Duration





Data Exploratory Analysis

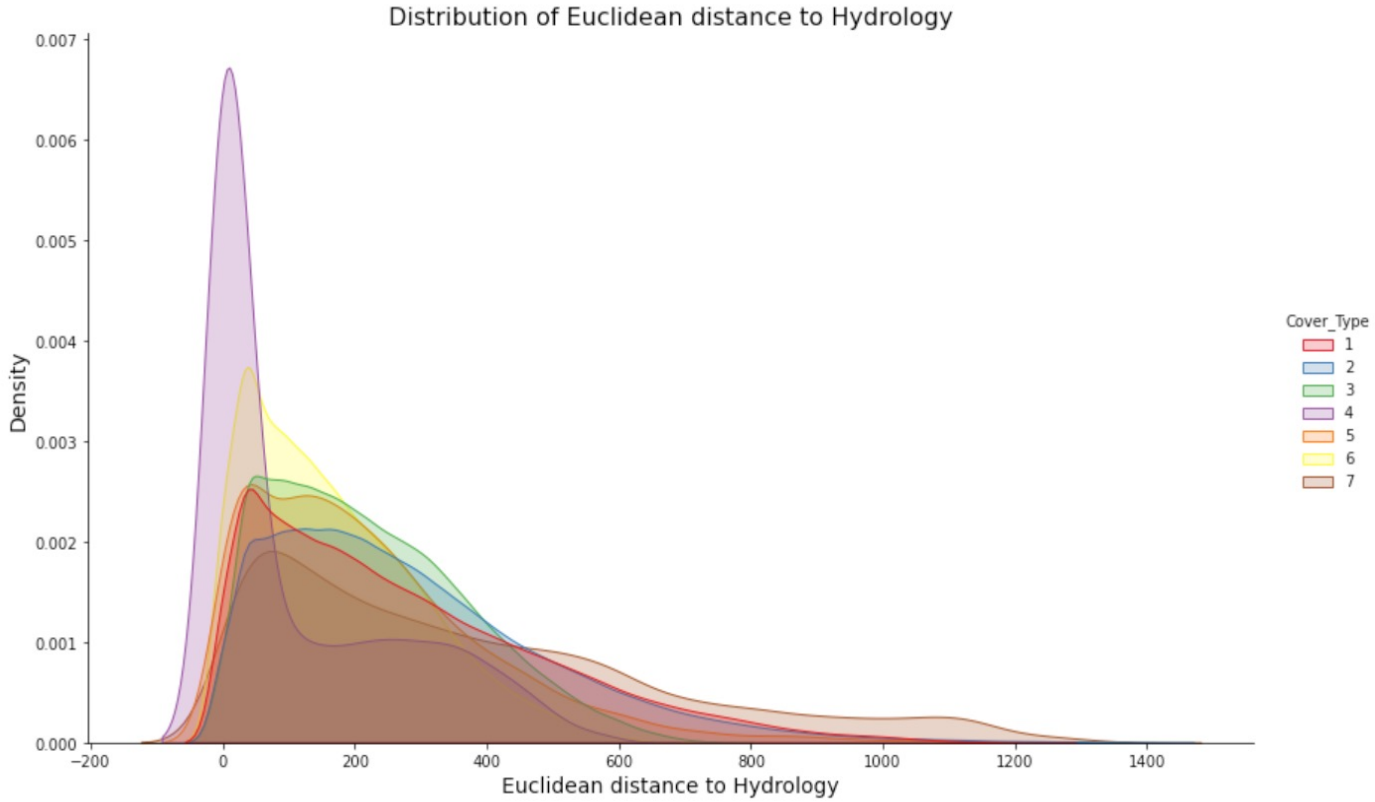
- Distribution of Elevation by Cover Type



- Strong variation on the distributions, indicating that elevation is a very useful predictor.
- Wider range: more adaptable to climate change, vice versa.

Data Exploratory Analysis

- Distribution of Euclidean Distance to Hydrology



- Based on the vertical and horizontal distance to hydrology, I created the euclidean distance to hydrology.
- All right skewed --> Thrive in moisture
- Cover type 4: needs more water or drought intolerant.

Model Selection

- Some Background

64% Training Set	16% Validation Set	20% Test Set

Metric used to measure the performance: Macro Average F1-score

- 4 Candidate Models

1. Decision Tree Model W/O Resampling
2. Decision Tree Model W/ Resampling
3. Random Forest Model W/O Resampling
4. Random Forest Model W/Resampling

The Resampling method we used is Borderline-SMOTE.

2	181312
1	135578
3	22882
7	13126
6	11115
5	6075
4	1759

Old Data(Training)

2	181312
1	135578
3	22882
7	13126
6	11115
5	10000
4	3000

New Data(Training)

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

Model 1: 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

Model 3: Random Forest W/O Resampling

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

Model 2: 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

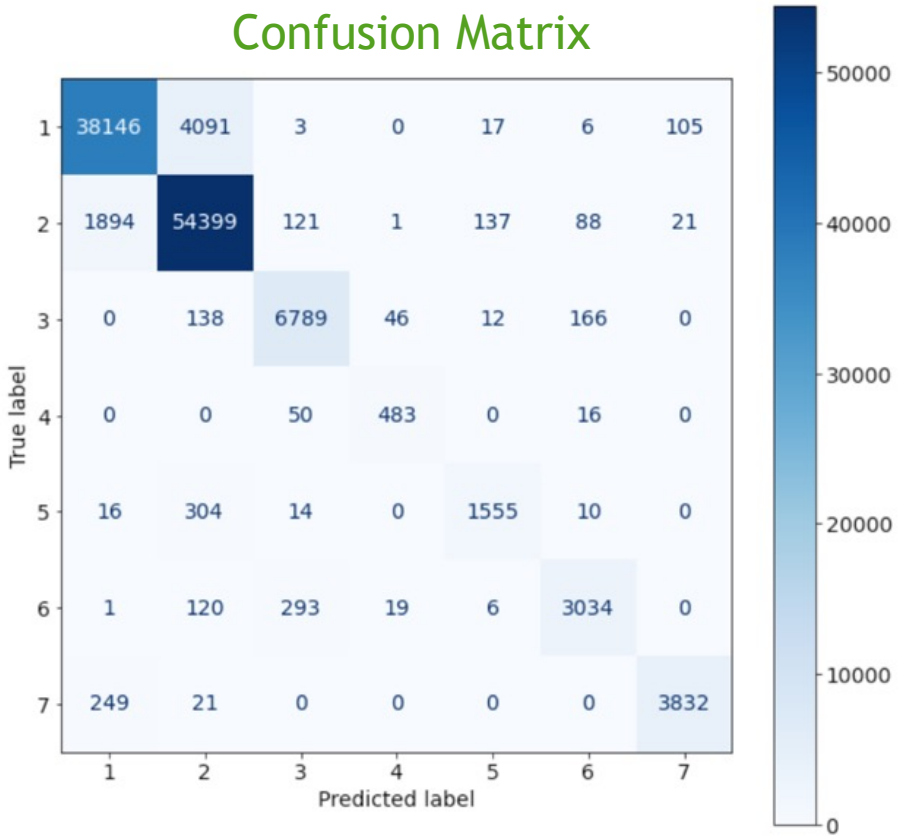
Model 4: Random Forest W/ Resampling

Results

Results on the test set

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

Confusion Matrix



The scores on the test set are close to the validation set!
Same macro avg f1-score and marco accuracy!
Performance on the minority classes 4 and 6 are stable.

Summary

- The main cover types in Different Wilderness Areas vary due to their local environmental conditions.
- The elevation matters the most.
- Cover types 3 and 2 are most adaptive to changing environment.
- Random Forest Model with resampling is the best model.

THANK YOU!