Forest Cover Type Prediction

Contents

1 Introduction

2 Data Pre-Processing

3 Different Classifiers

4 Hierarchical Method

5 Feature Engineering

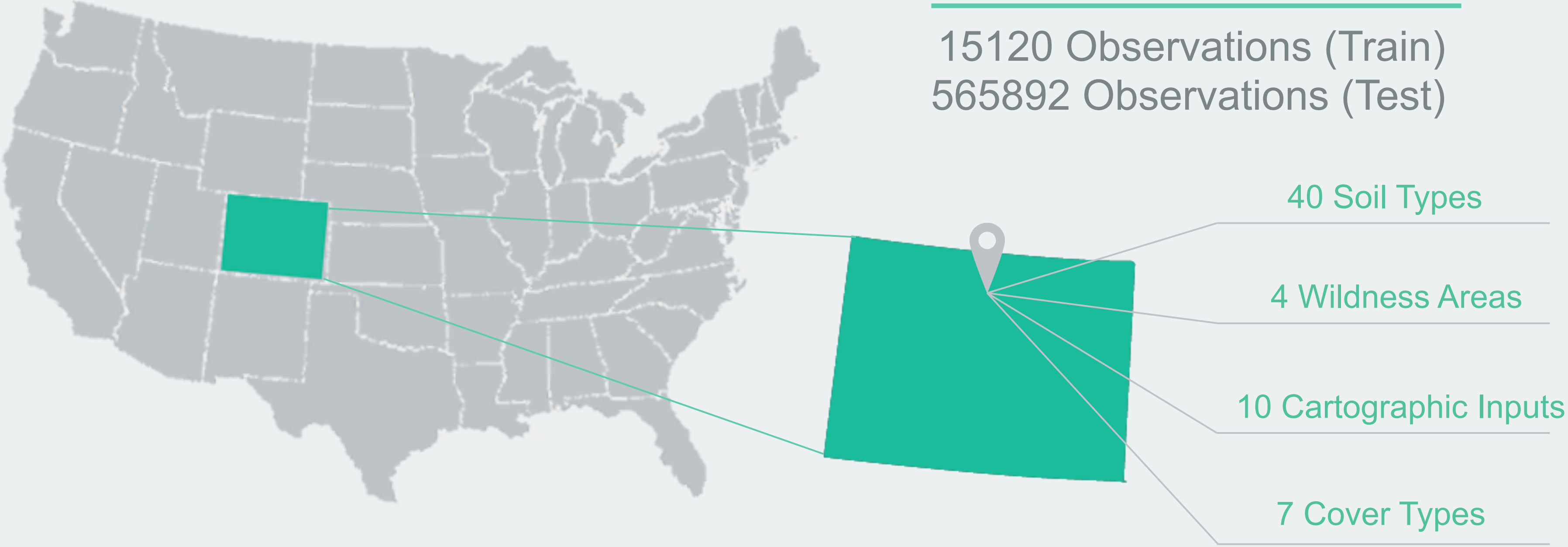
6 Post-Model Analysis

1 Introduction

Introduction

Roosevelt National Forest

15120 Observations (Train)
565892 Observations (Test)



2 Data Pre-Processing

Pre-Processing

Quick Visualization

Get a general idea of variable type, mean, range, etc.

Data Cleaning

Deal with missing values.

Feature Extraction

Extract features from original attributes.

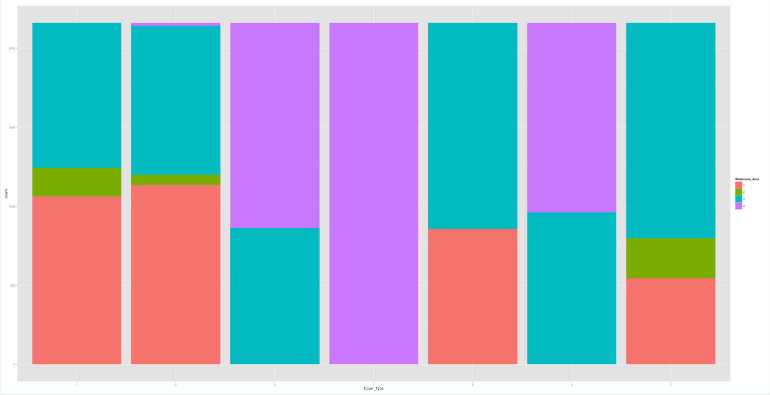
Validation Subset

Randomly sample 1/3 of training data as validation set.

Quick Visualization

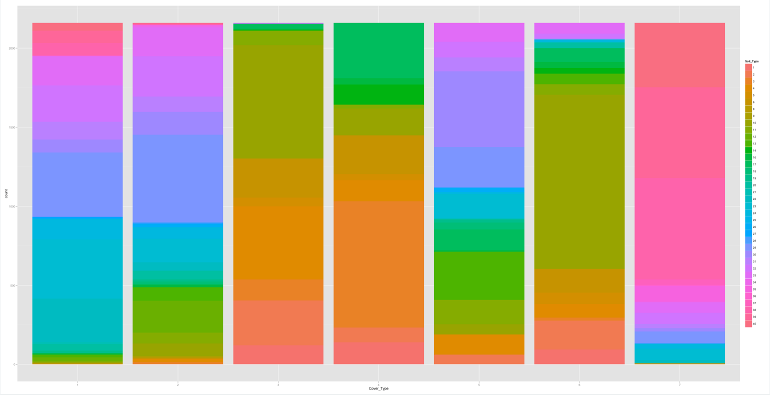
WA Distribution

Wilderness area distribution of different cover types

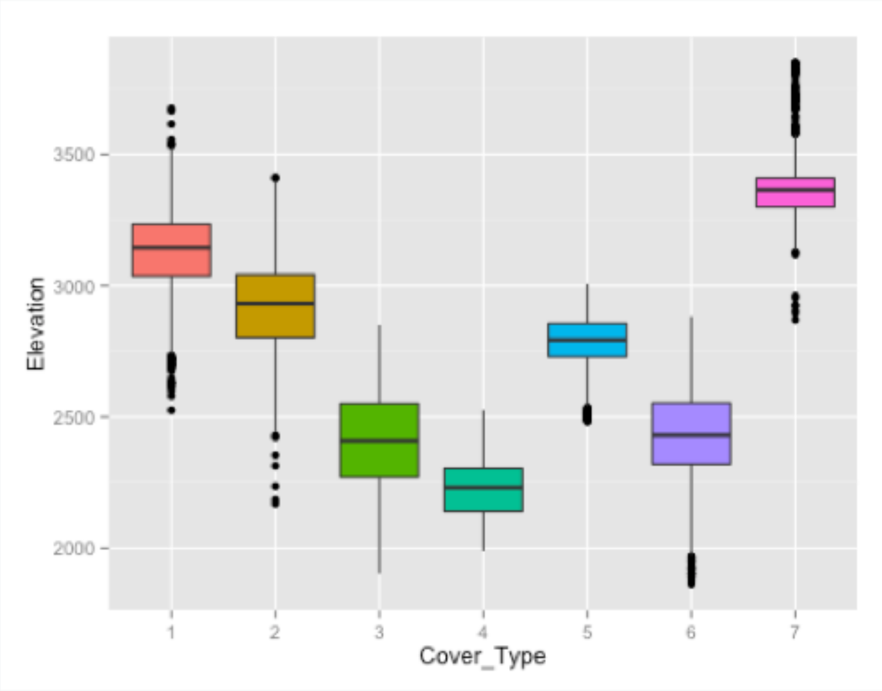


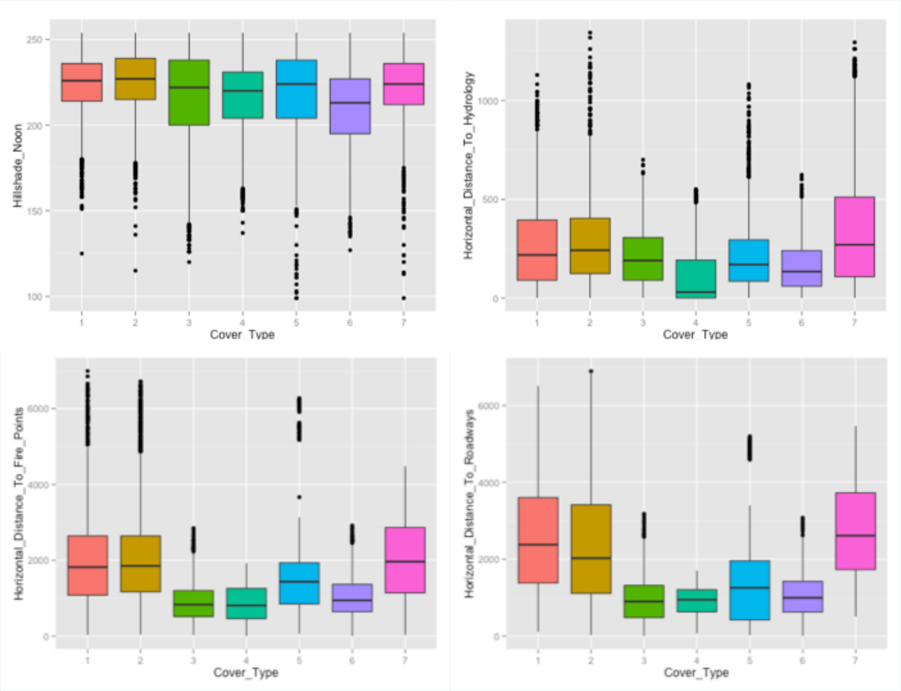
ST Distribution

Soil type distribution of different cover types

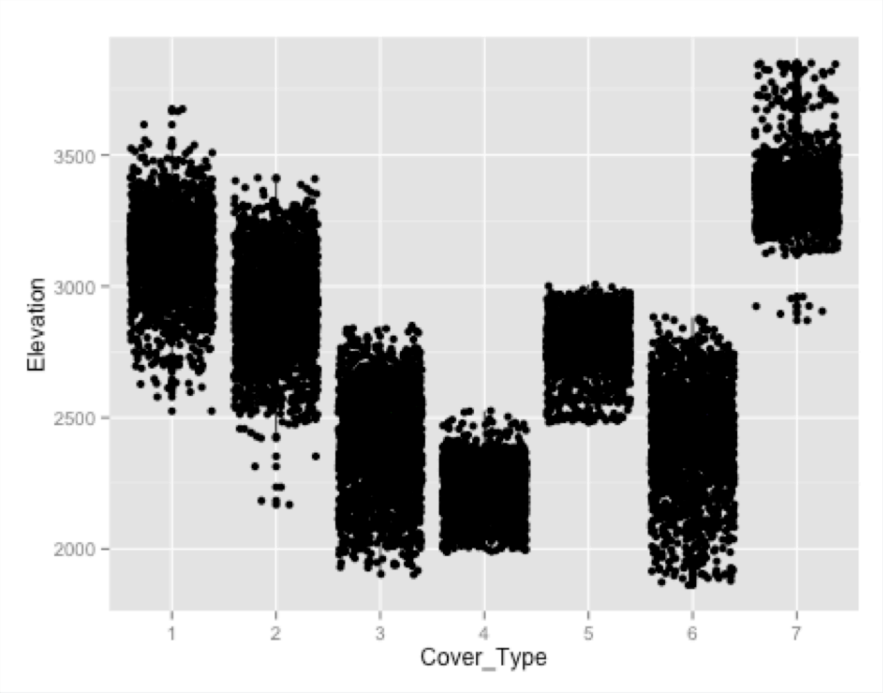


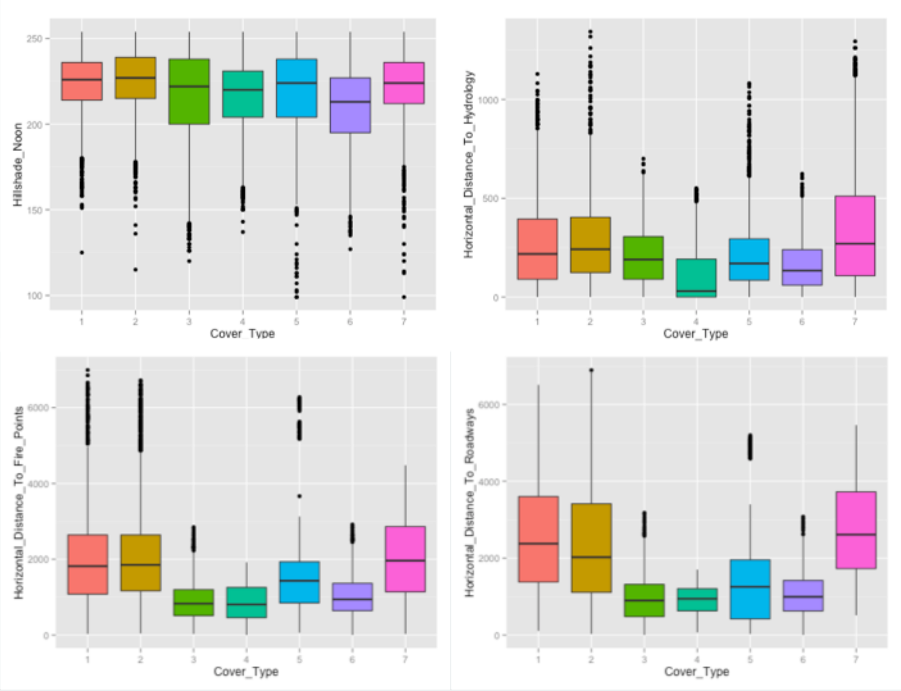
Quick Visualization





Quick Visualization





Data Cleaning

Soil Type Missing

Missing in training but existing in test

USFS ELU Code

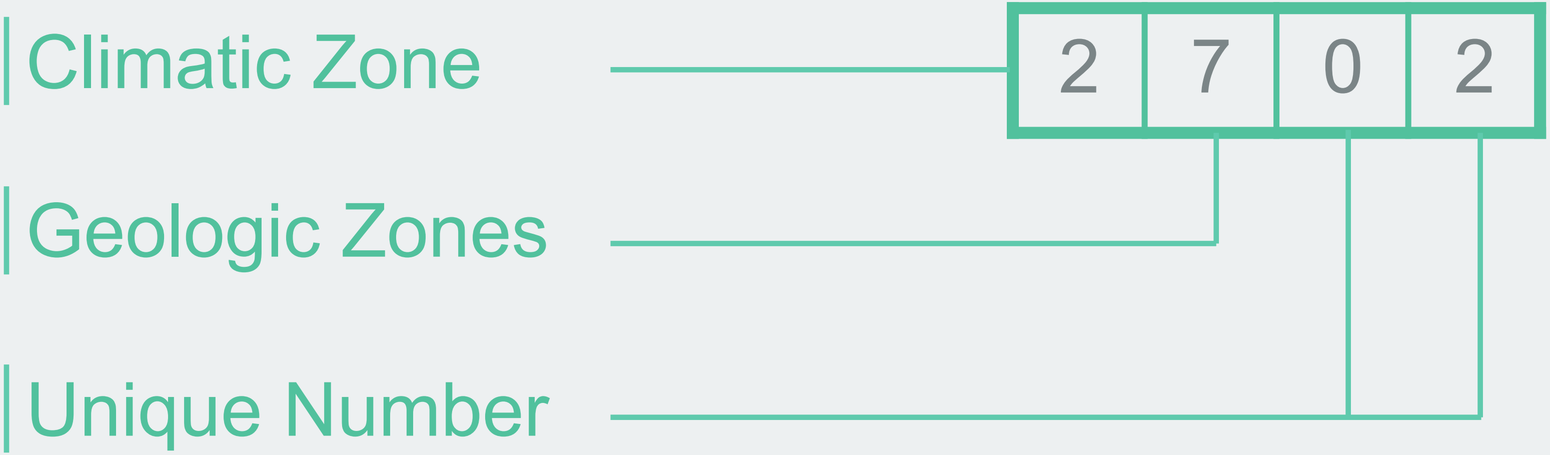
USFS (United States Forest Service) Ecological Landscape Units

Climatic & Geologic Zone

8 climatic zones and 8 geologic zones

Feature Extraction

USFS ELU Code



2	7	0	2

Climatic Zone

Geologic Zones

Unique Number

Feature Extraction

Method

Common Sense

Physical Equation

Geologic Knowledge

Hillshade Variations

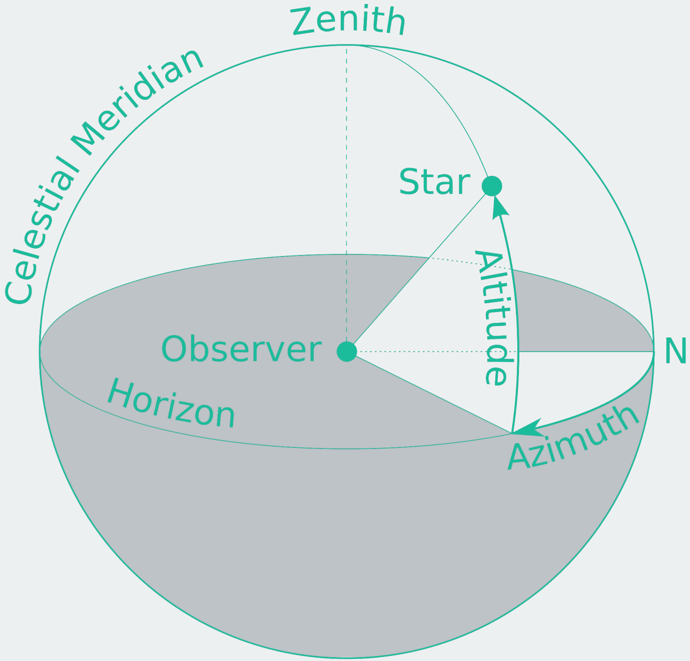
Temperature

Data

Geologic Zone

Climatic Zone

Feature Extraction



Hillshade Algorithm

$\text{Hillshade} = 255.0\cdot(\cos(Z)\cdot\cos(S)+\sin(Z)\cdot\sin(S)\cdot\cos(Az\text{-}As)\ )$

Z: Zenith S: Slope Az: Azimuth As: Aspect

$\text{Hillshade Variation} =1/3\cdot\Sigma\ [\text{Hillshade}(i)\text{– Hillshade(mean)}]^2$

Temperature-Altitude Equation

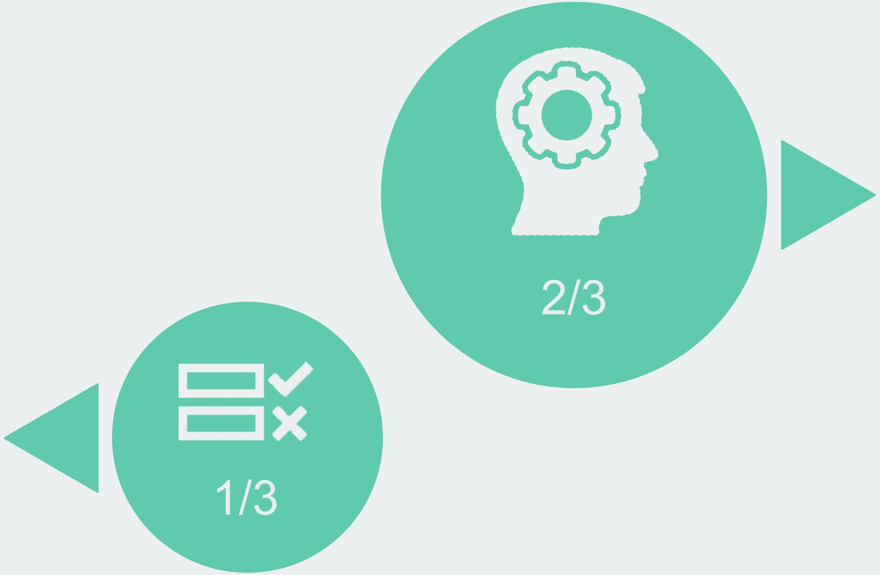
$T = T_0 - 6.5\cdot(H/1000)$

Validation Subset

Ratio

10080 for training

5040 for validation

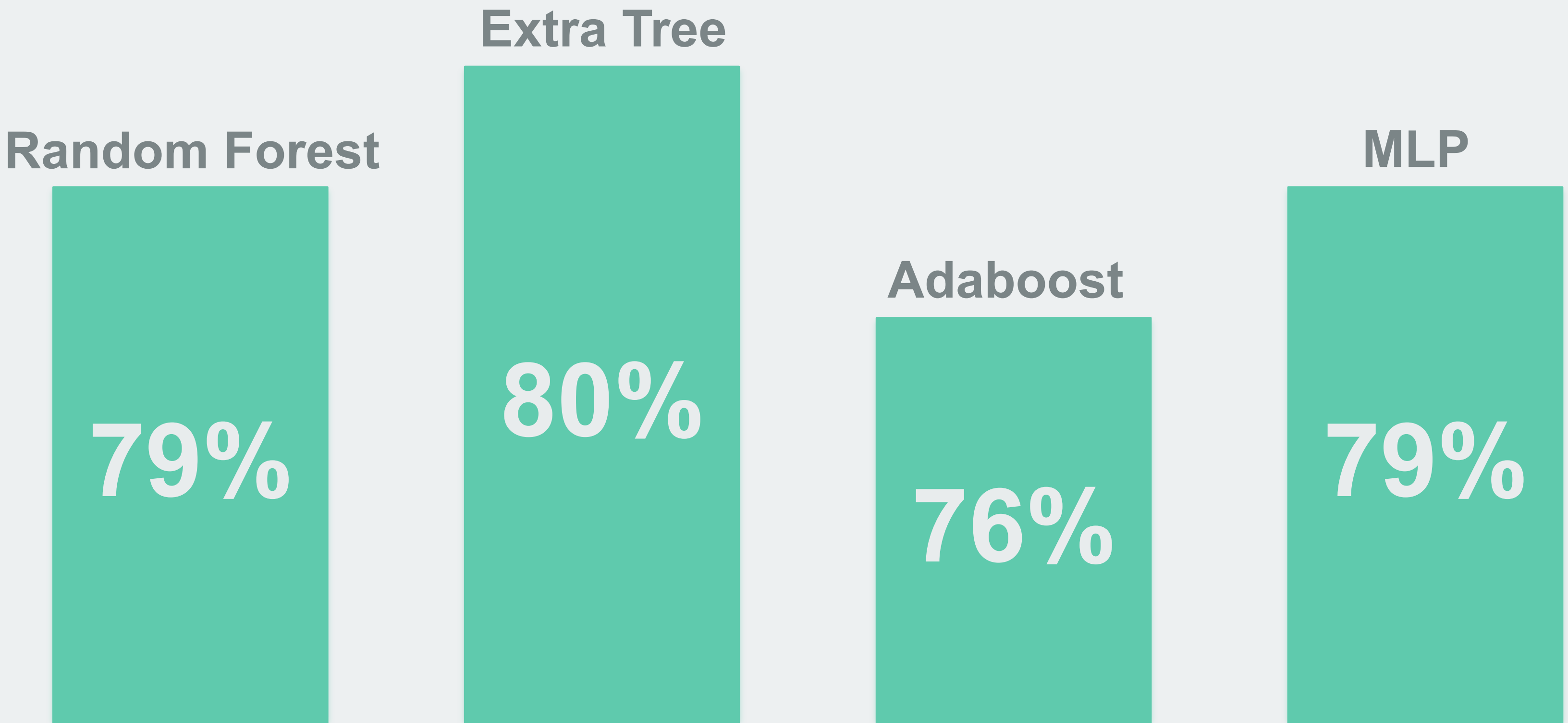


2/3

1/3

3 Different Classifiers

Accuracy of Different Classifiers



4 Hierarchical Method

Confusion Matrix

		Reference						
		1	2	3	4	5	6	7
Prediction	1	473	165	0	0	10	0	49
	2	158	408	14	0	59	9	10
	3	1	13	532	29	16	120	0
	4	0	0	36	639	0	29	0
	5	22	100	17	0	650	9	1
	6	4	19	117	13	10	543	0
	7	74	11	0	0	0	0	680

Prediction Result

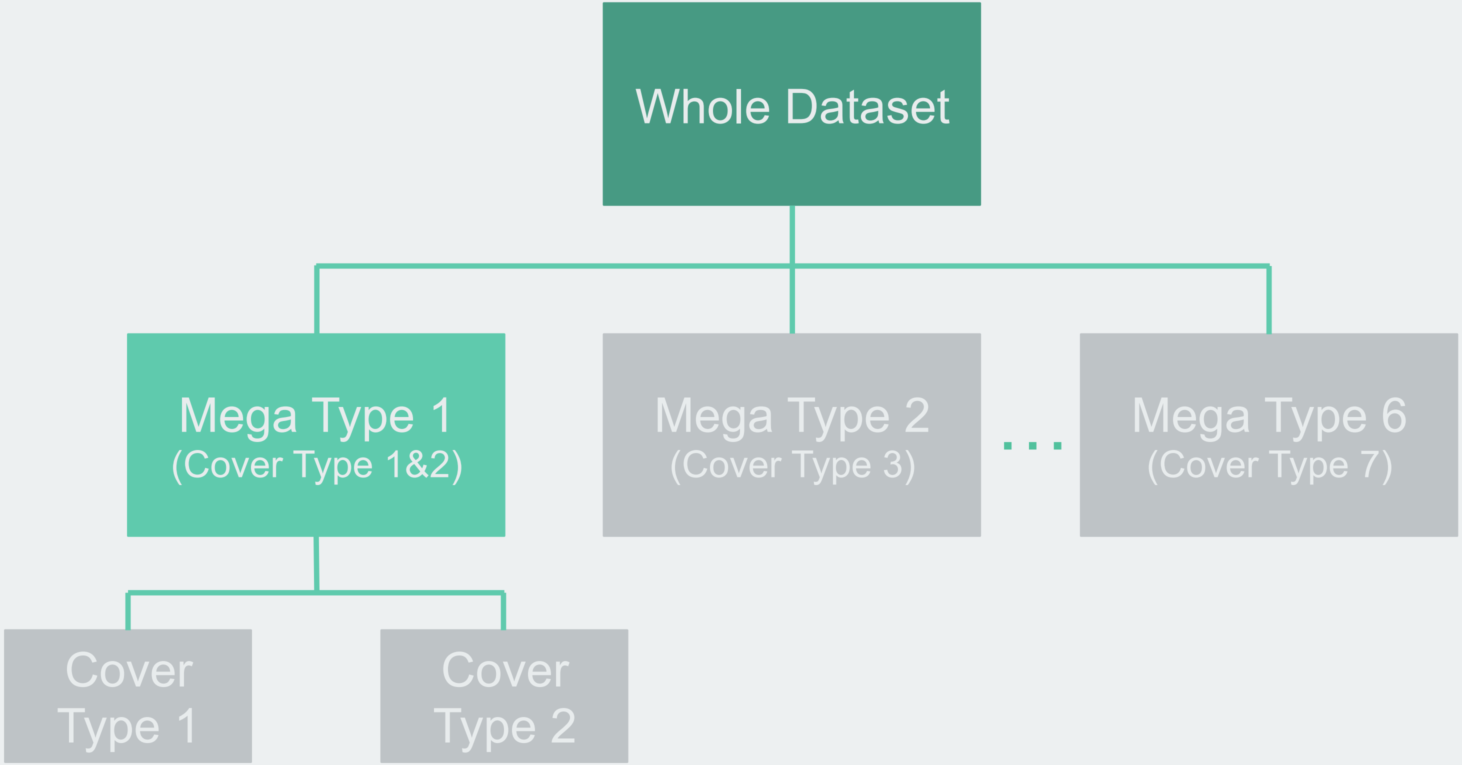
(on validation subset)

Cover Type 1 & Cover Type 2 are easily confused with each other.

The test result shows that in test dataset Cover Type 1 & 2 predominate over other types.

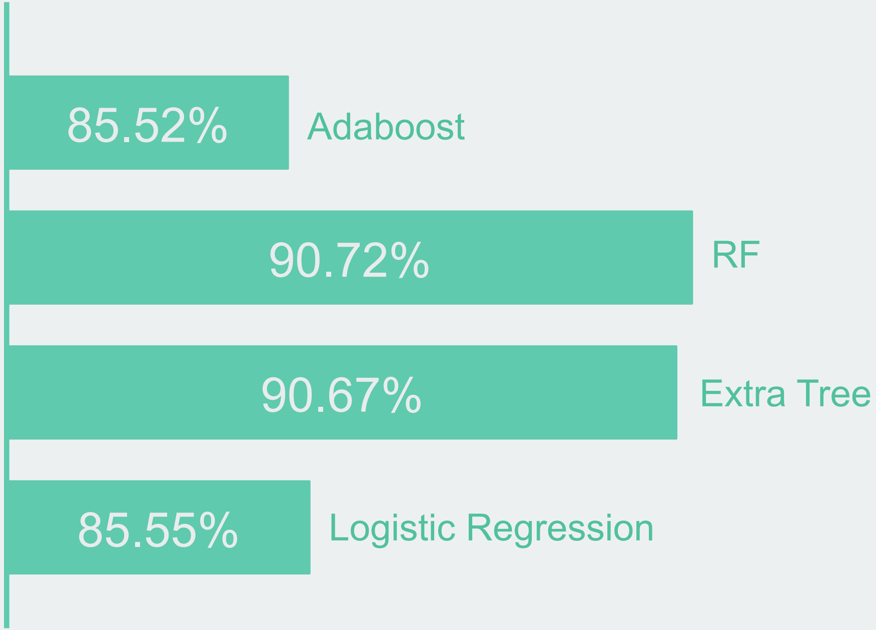
Combine Cover Type 1 & 2 as a mega type!

Two-Level Hierarchical Classifier

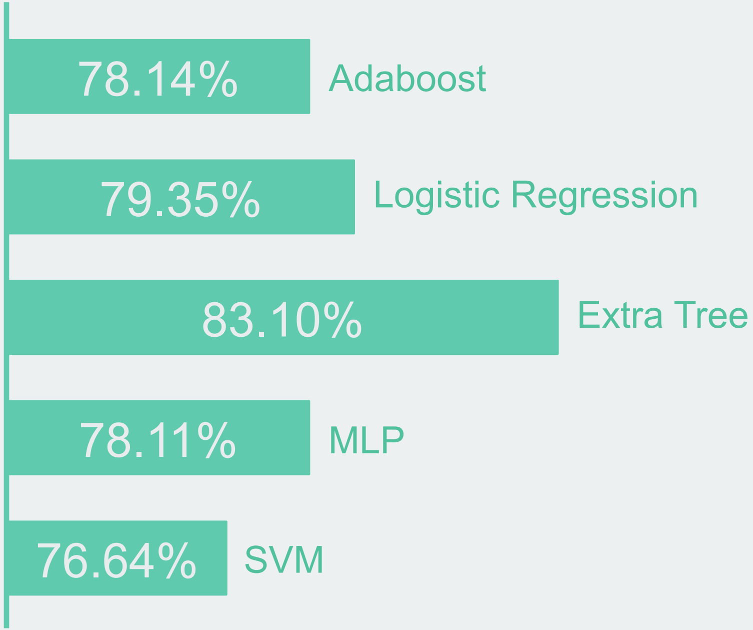


Accuracy of Each Level

First Level



Second Level



5 Feature Engineering

Feature Engineering

1 Finding features that better represent the underlying problem to the predictive model

2 Discarding features likely to expose us to the risk of over-fitting

3 Trying different encodings for wilderness area and soil type
→ will not help

Feature Engineering

New Features

EVDtH, EHDtH, Fire_Road_1, Hydro_Fire_2

Improvement

3% improvement on average

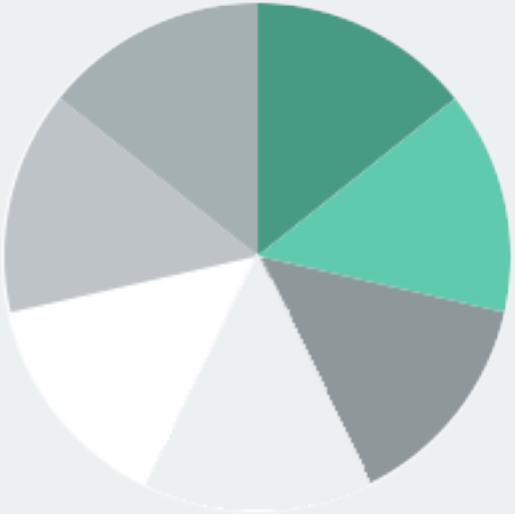
Features	Importance
EHDtH	0.097486
Elevation	0.096896
EVDtH	0.092564
Wilderness_Area4	0.046237
Fire_Road_1	0.033677
Hydro_Road_2	0.032912
Horizontal_Distance_to_Roadways	0.031300
Hydro_Road_1	0.030773
Distance_to Hydrology	0.028715
Horizontal_Distance_to_Hydrology	0.027839

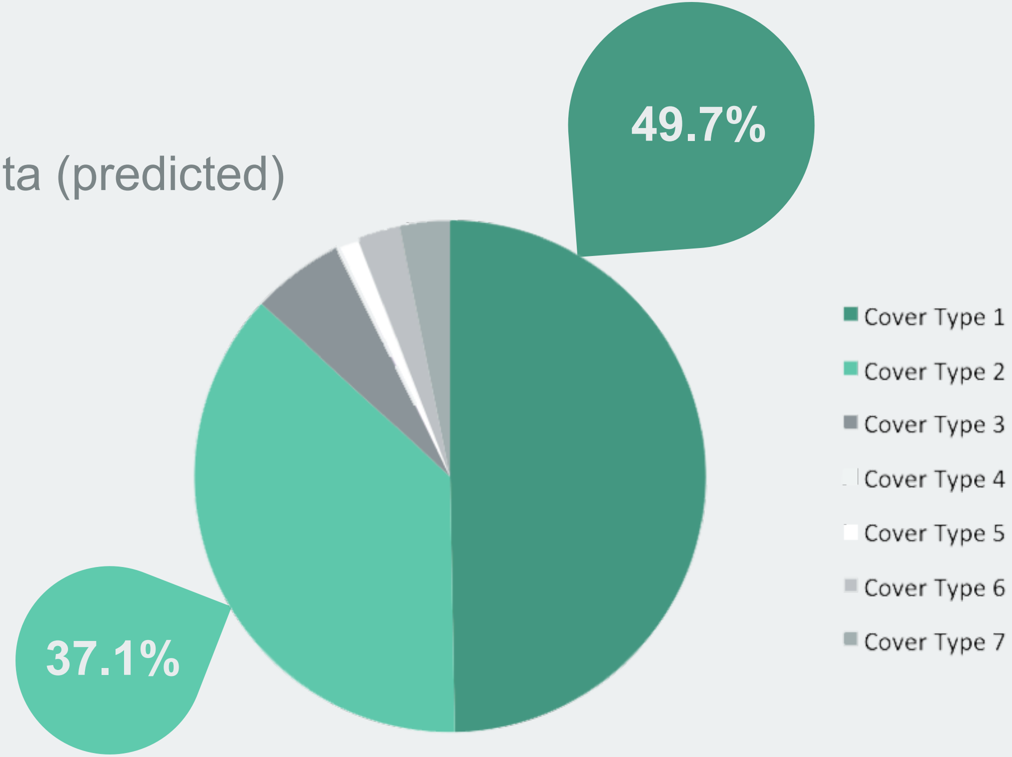
6 Post-Model Analysis

Change Prior

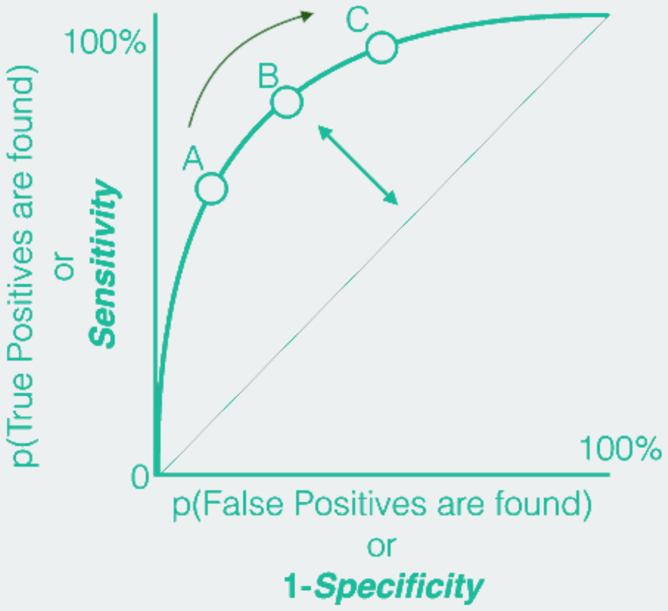
Prior Comparison

Priors of training data and test data (predicted)





Resampling



$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$= \frac{\text{Number of Predicted Class } 1 \& 2}{\text{Number of Actual Class } 1 \& 2}$$

Final Result

Kaggle Ranking

10

84.472% ✓

15.528% ✗

THANKS