The Amazon from Space: Machine Learning applied to deforestation

NICK McCORMACK
ELEANOR PROUST
WEI WANG
SAI RAVURU

8/9/2018

Agenda

- Problem statement
- Process
- EDA & Correlation
- Image Processing
- Weather feature prediction
- Land feature prediction
- Conclusion

Problem/Data Context

Our Goals

□ The rainforest disappears at a rate equal to several football fields a day

□ We want to build a model that accurately identifies when forest disappears

□ Satellites take photos of land all the time, meaning this is a problem machine learning can solve

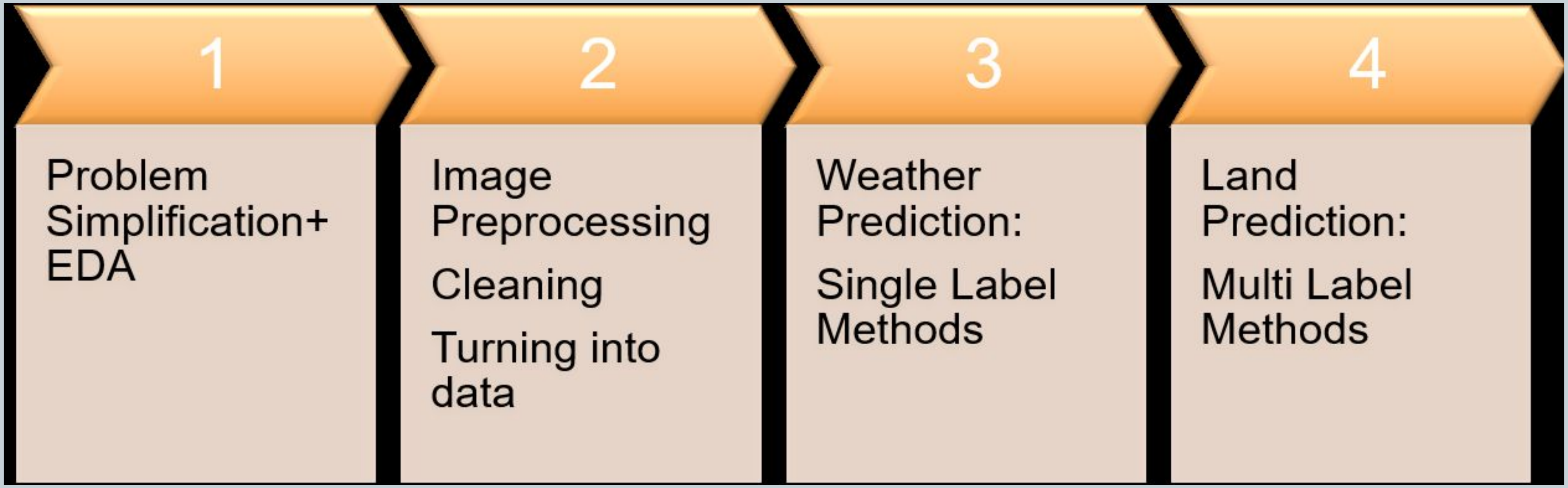
Problem/Data Context

- 40K Images - two types standard JPGs and infrared data
- Each Image Contains multiple labels - one weather and any or none from 11 different land labels

Tag 1	Tag 2	Tag 3	Tag 4	Tag 5
agriculture	clear	habitation	primary	road
agriculture	clear	primary	water	
haze	primary			

- Given we care about the destruction of nature – we care about identifying natural features, and man made features

Our Process



1

Problem Simplification+ EDA

2

Image Preprocessing

Cleaning

Turning into data

3

Weather Prediction:

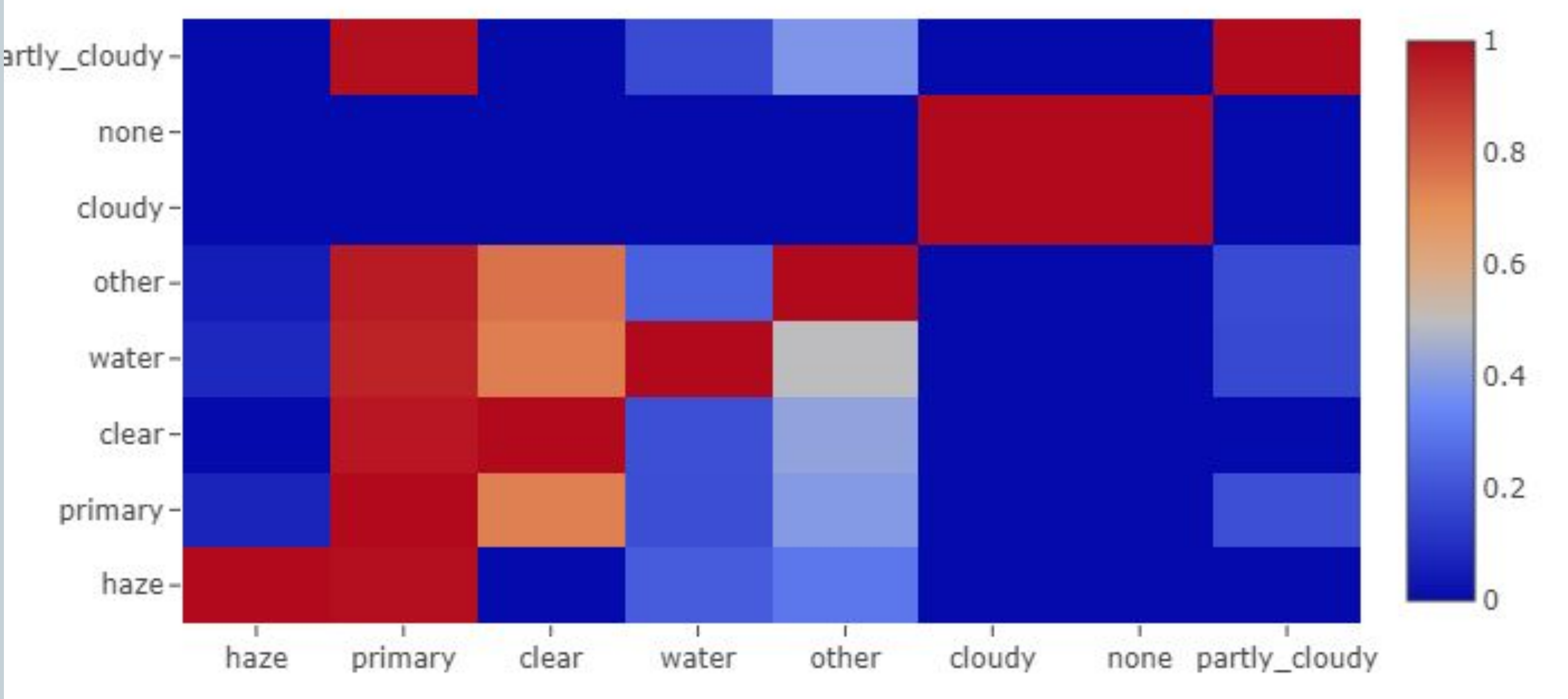
Single Label Methods

4

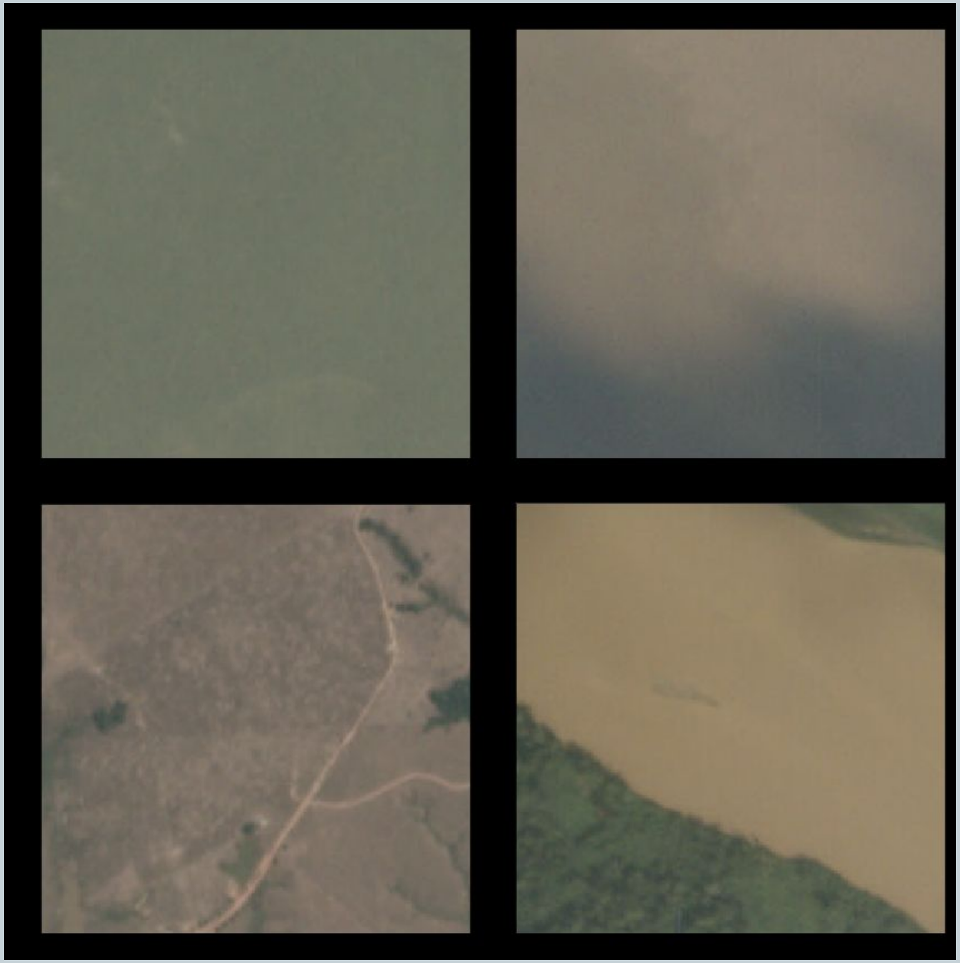
Land Prediction:

Multi Label Methods

EDA - Correlation between labels



Example of the problem



Hazy
Primary

Primary
Clear

Clear
Other

Primary
Clear
Water

Image Processing - Haze Removal

- Atmospheric light intensity is measured.
- OpenCV filters used to remove cloud and haze.
- Underlying features are highlighted to foreground.

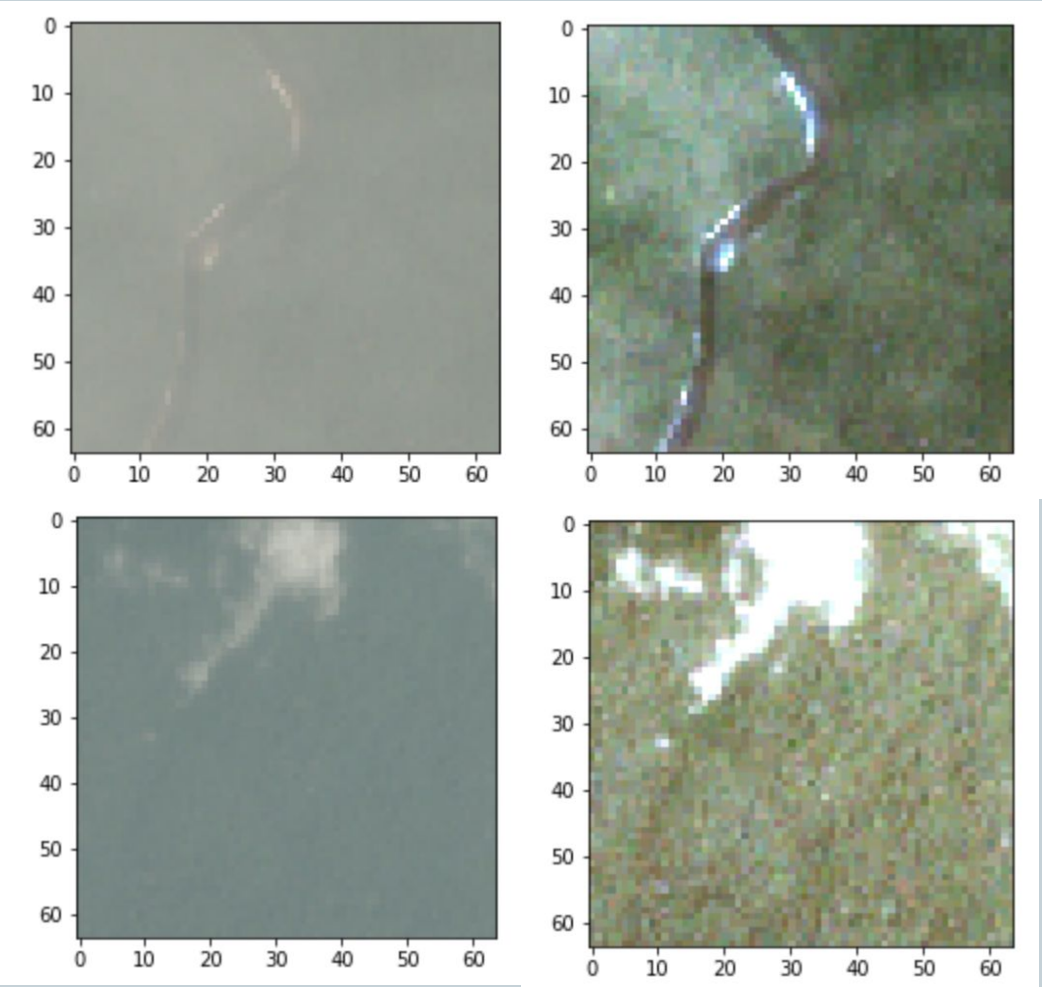
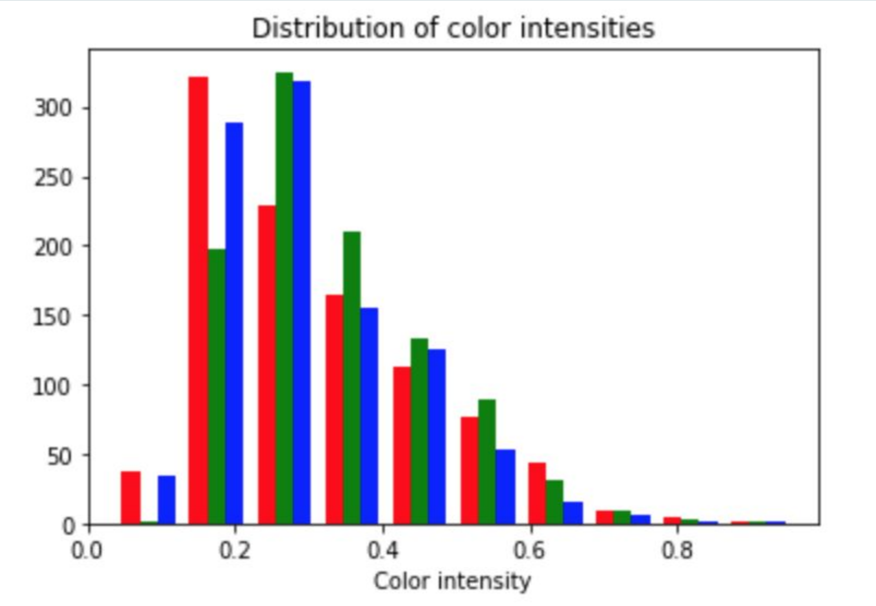
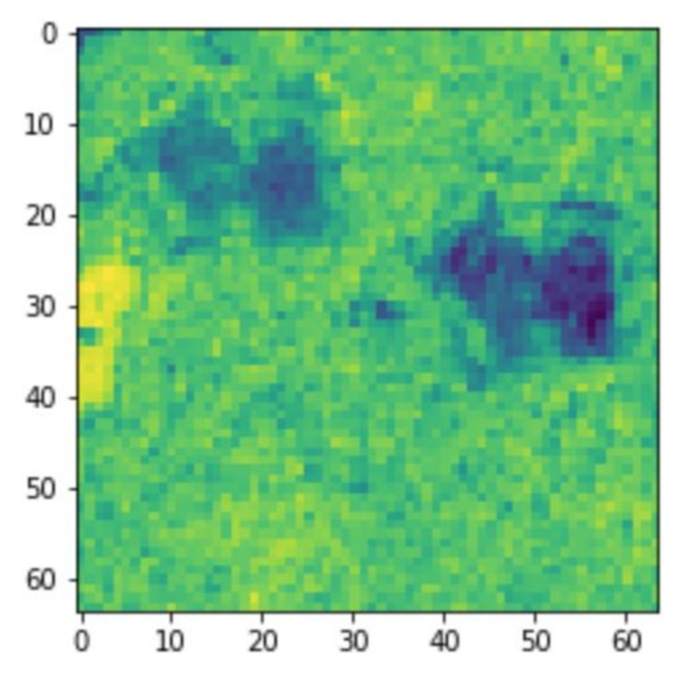


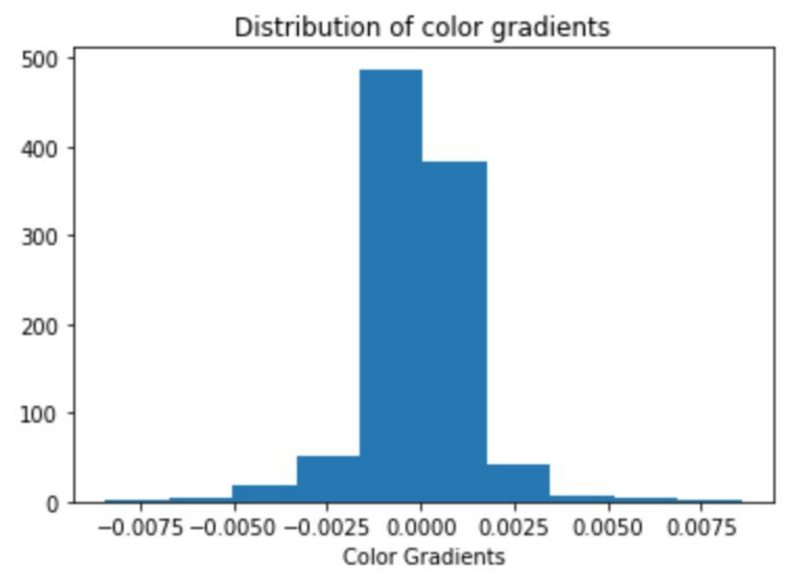
Image Processing - Spectral Analysis

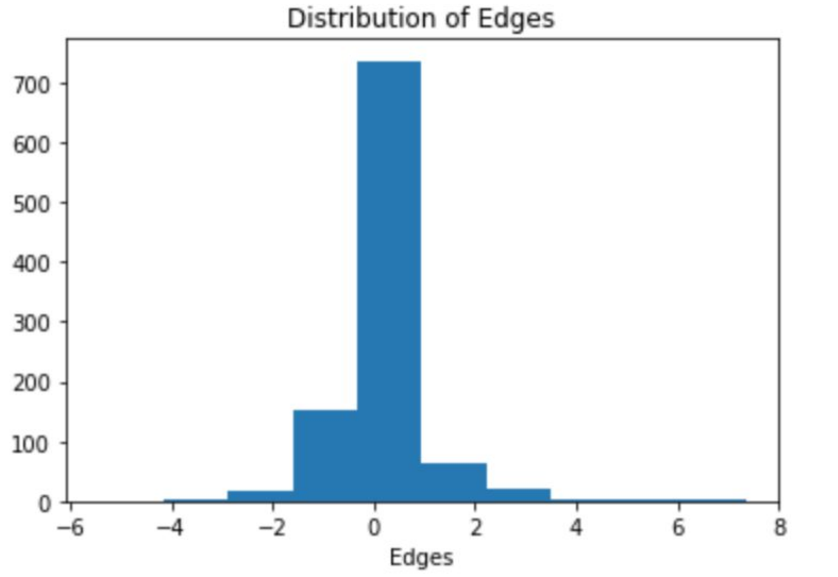


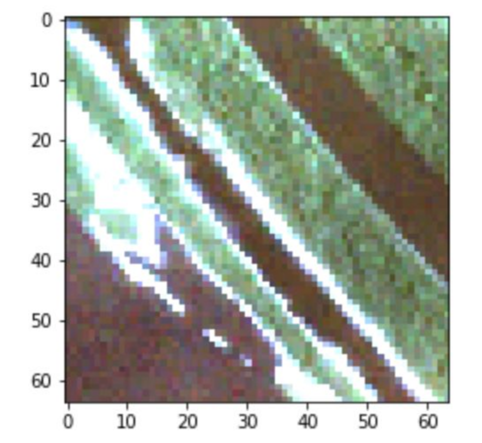


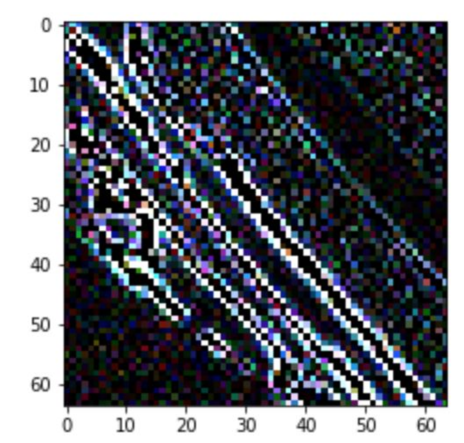
- RGB intensity and distribution are low due to image resolution.
- Blue and Green dominance indicate high volume of water & vegetation.

Image Processing - Edge Detection


Distribution of color gradients


Distribution of Edges





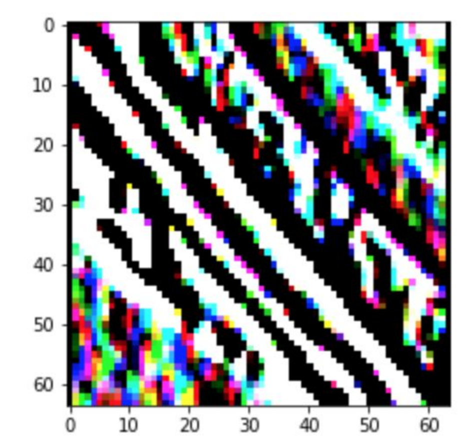
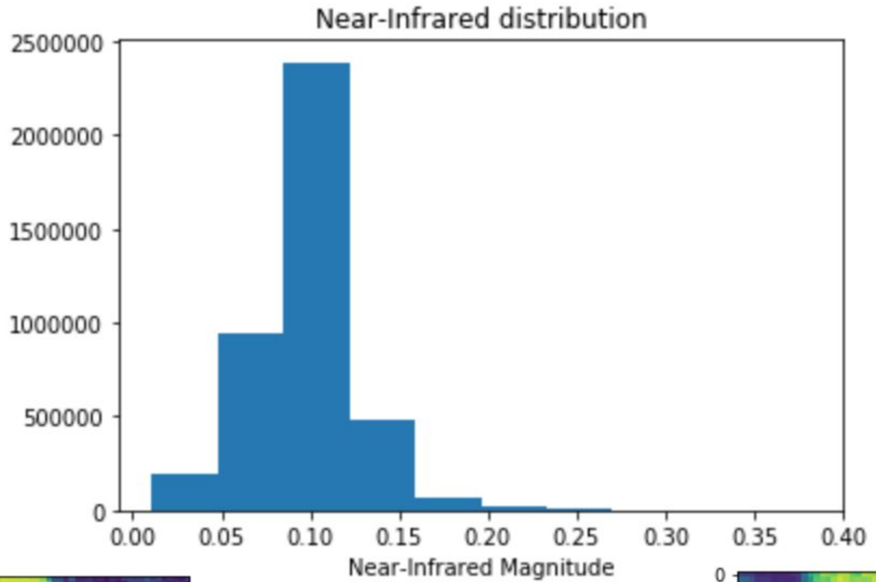
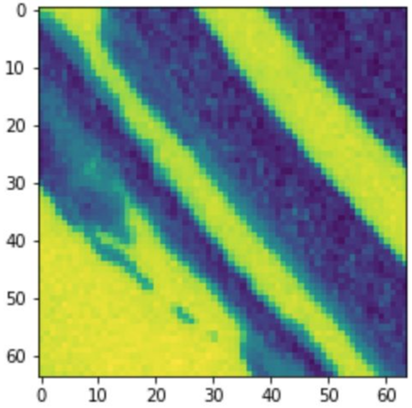


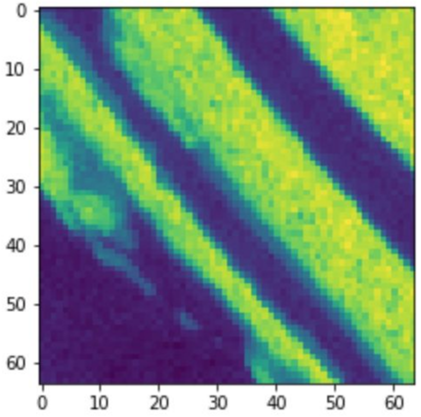
Image Processing - Near-Infrared

Blue frequency (Water): (B-IRR)/(B+IRR)

Green frequency (Vegetation): (G-IRR)/(G+IRR)





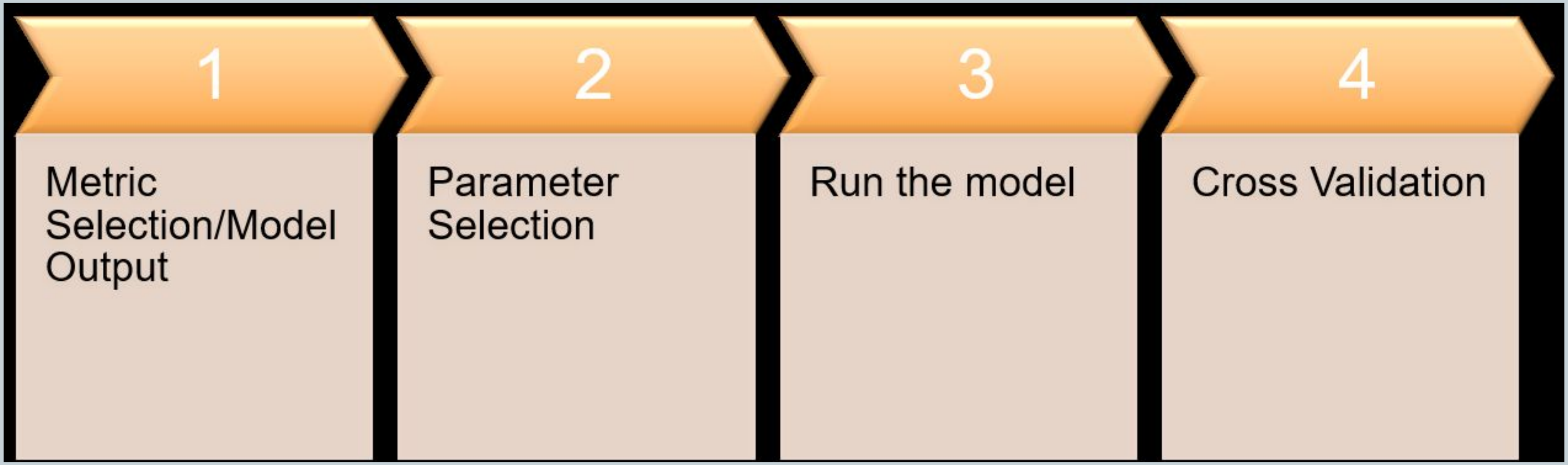


Weather Prediction Models Overview

Purpose:

- Satellite data is not always clear, a lack of forest may just be cloud
- We don’t want to predict deforestation when we just can’t get a clear photo!

Our Pipeline:



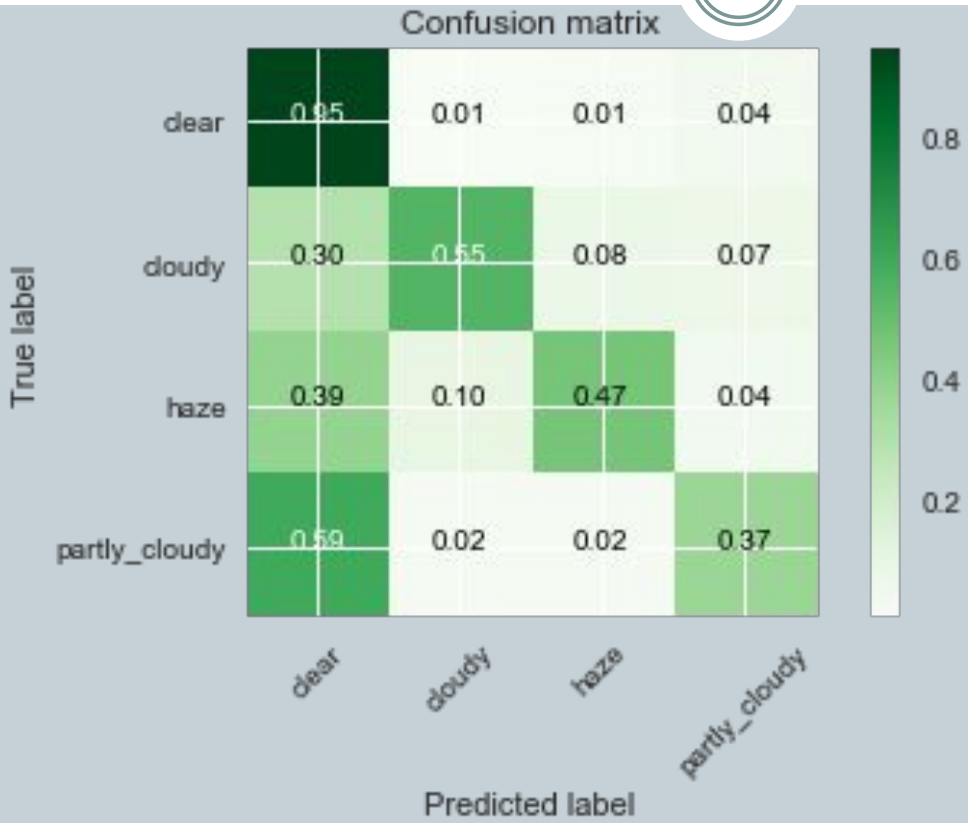
1 Metric Selection/Model Output

2 Parameter Selection

3 Run the model

4 Cross Validation

Weather Prediction - Logistic Regression

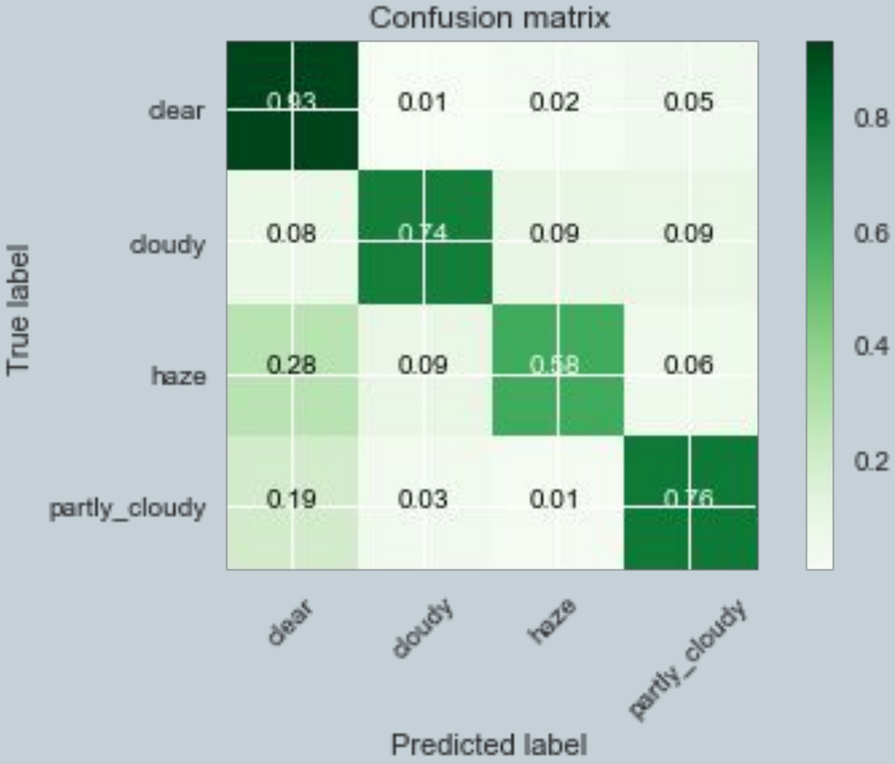


Cross-Validation Score

Accuracy: 0.77 (+/- 0.01)

- The result shows that the "Cloudy" and "Haze" are predicted very poorly whereas the "Clear" and "Partly_cloudy" are much more accurate.

Weather Prediction - Decision Tree

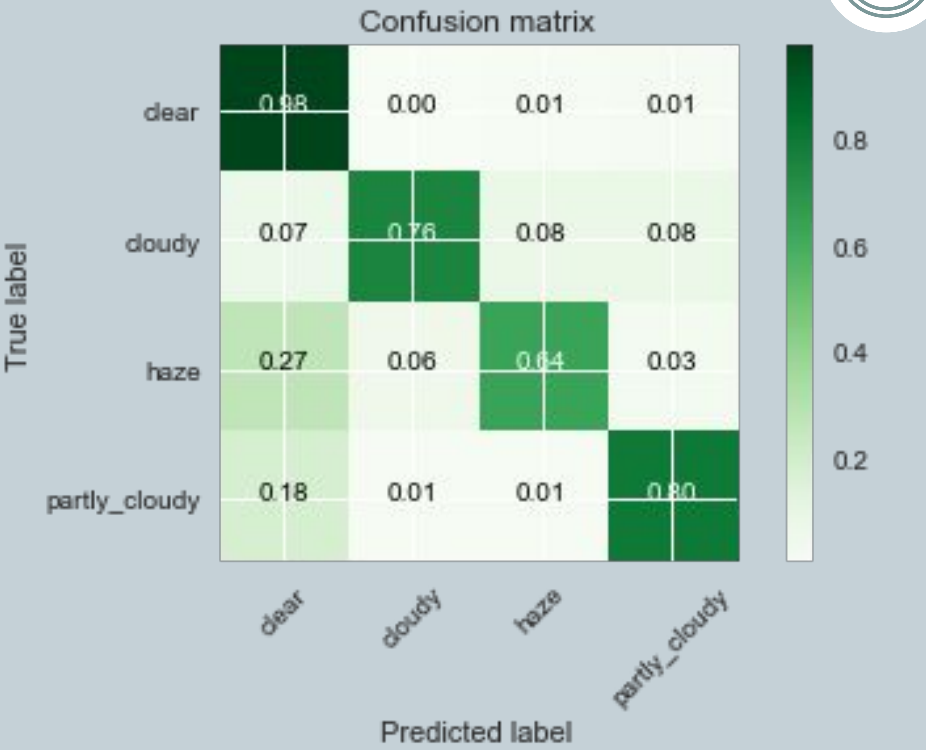


Cross-Validation Score

```
Accuracy: 0.87 (+/- 0.01)
```

- Decision tree model greatly improves the accuracy on all of the rare tags
- It also has better cross validation score
- Thus it's a better model than Logistics Regression

Weather Prediction - Random Forest



Cross-Validation Score

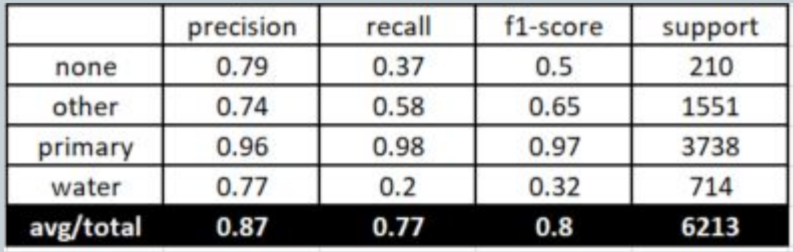
Accuracy: 0.90 (+/- 0.01)

- Compared to decision tree, Random Forest model greatly improves the scores for “Cloudy” and “Clear”
- Cloudy gets confused for all other classes equally - though hazy/partly cloudy is confused often for “Clear”)
- It also performs better for cross-validation - this is our best weather model

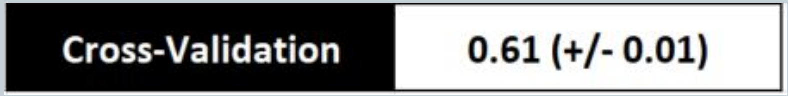
Land Prediction - K-Nearest Neighbors

- Optimal performance with k = 12
- Trouble identifying water
- Identified "other" tags as "primary" almost 40% of the time when classified
- f1-score and cross-validation score were lowest of the multi-label classification models

Classification Report

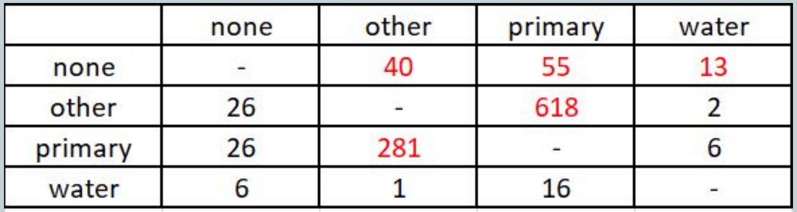


	precision	recall	f1-score	support
none	0.79	0.37	0.5	210
other	0.74	0.58	0.65	1551
primary	0.96	0.98	0.97	3738
water	0.77	0.2	0.32	714
avg/total	0.87	0.77	0.8	6213

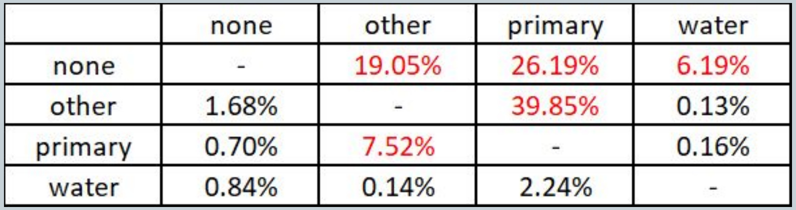


Cross-Validation	0.61 (+/- 0.01)

Confusion Matrices



	none	other	primary	water
none	-	40	55	13
other	26	-	618	2
primary	26	281	-	6
water	6	1	16	-



	none	other	primary	water
none	-	19.05%	26.19%	6.19%
other	1.68%	-	39.85%	0.13%
primary	0.70%	7.52%	-	0.16%
water	0.84%	0.14%	2.24%	-

Land Models - One vs. the Rest (OvR)

Random Forest

vs

Classification Report

	precision	recall	f1-score	support
none	0.83	0.73	0.77	210
other	0.85	0.82	0.83	1551
primary	0.97	0.99	0.98	3738
water	0.82	0.59	0.69	714
avg/total	0.92	0.89	0.9	6213

Cross-Validation	0.77 (+/- 0.00)

Confusion Matrices

	none	other	primary	water
none	-	7	38	5
other	9	-	273	5
primary	26	207	-	3
water	14	2	6	-

	none	other	primary	water
none	-	3.33%	18.10%	2.38%
other	0.58%	-	17.60%	0.32%
primary	0.70%	5.54%	-	0.08%
water	1.96%	0.28%	0.84%	-

Decision Tree Estimator

Classification Report

	precision	recall	f1-score	support
none	0.7	0.69	0.7	210
other	0.8	0.78	0.79	1551
primary	0.97	0.98	0.98	3738
water	0.64	0.57	0.61	714
avg/total	0.88	0.87	0.88	6213

Cross-Validation	0.69 (+/- 0.01)

Confusion Matrices

	none	other	primary	water
none	-	11	24	7
other	24	-	323	2
primary	47	279	-	5
water	17	2	6	-

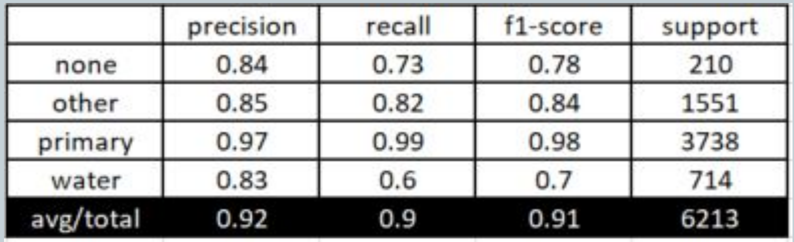
	none	other	primary	water
none	-	5.24%	11.43%	3.33%
other	1.55%	-	20.83%	0.13%
primary	1.26%	7.46%	-	0.13%
water	2.38%	0.28%	0.84%	-

Land Models -Classifier Chains

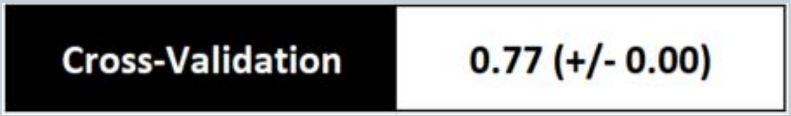
Random Forest

vs

Classification Report

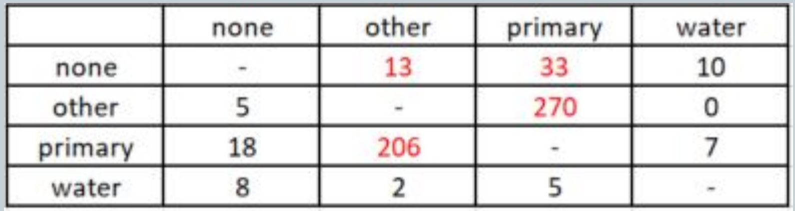


	precision	recall	f1-score	support
none	0.84	0.73	0.78	210
other	0.85	0.82	0.84	1551
primary	0.97	0.99	0.98	3738
water	0.83	0.6	0.7	714
avg/total	0.92	0.9	0.91	6213

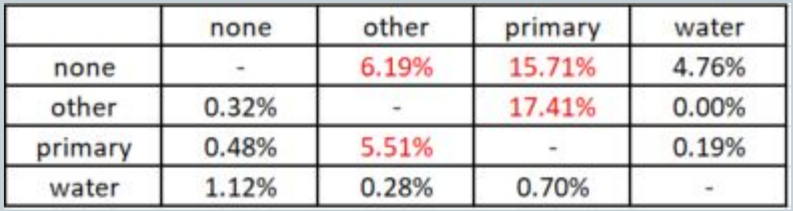


Cross-Validation	0.77 (+/- 0.00)

Confusion Matrices



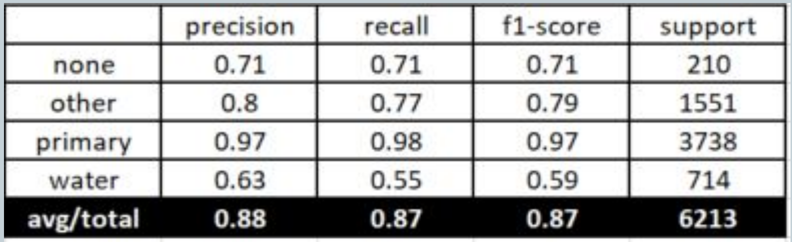
	none	other	primary	water
none	-	13	33	10
other	5	-	270	0
primary	18	206	-	7
water	8	2	5	-



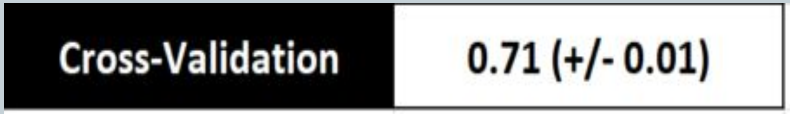
	none	other	primary	water
none	-	6.19%	15.71%	4.76%
other	0.32%	-	17.41%	0.00%
primary	0.48%	5.51%	-	0.19%
water	1.12%	0.28%	0.70%	-

Decision Tree Estimator

Classification Report

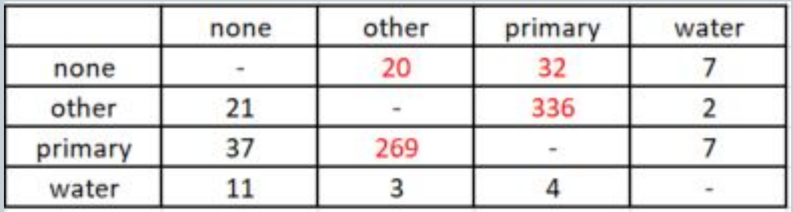


	precision	recall	f1-score	support
none	0.71	0.71	0.71	210
other	0.8	0.77	0.79	1551
primary	0.97	0.98	0.97	3738
water	0.63	0.55	0.59	714
avg/total	0.88	0.87	0.87	6213

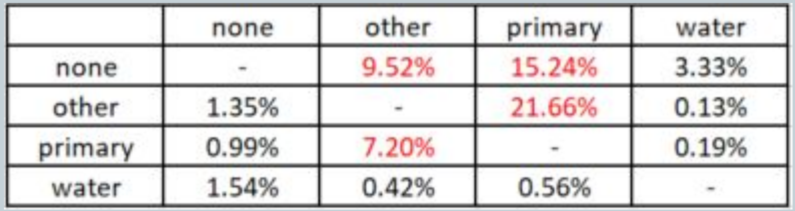


Cross-Validation	0.71 (+/- 0.01)

Confusion Matrices



	none	other	primary	water
none	-	20	32	7
other	21	-	336	2
primary	37	269	-	7
water	11	3	4	-



	none	other	primary	water
none	-	9.52%	15.24%	3.33%
other	1.35%	-	21.66%	0.13%
primary	0.99%	7.20%	-	0.19%
water	1.54%	0.42%	0.56%	-

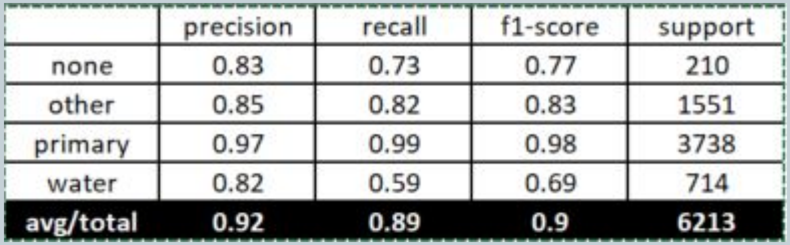
Land Models - Selection of Multilabel Model with Random Forest Estimator

OvR

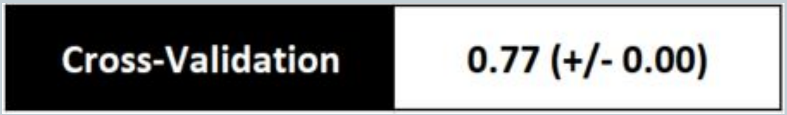
vs

Classifier Chain

Classification Report

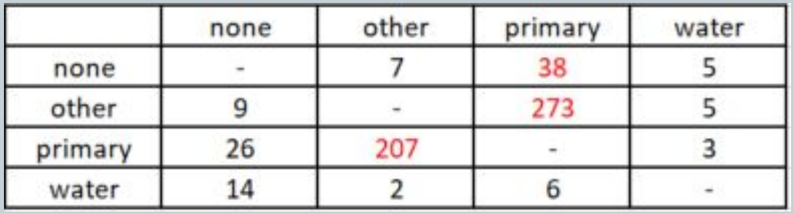


	precision	recall	f1-score	support
none	0.83	0.73	0.77	210
other	0.85	0.82	0.83	1551
primary	0.97	0.99	0.98	3738
water	0.82	0.59	0.69	714
avg/total	0.92	0.89	0.9	6213

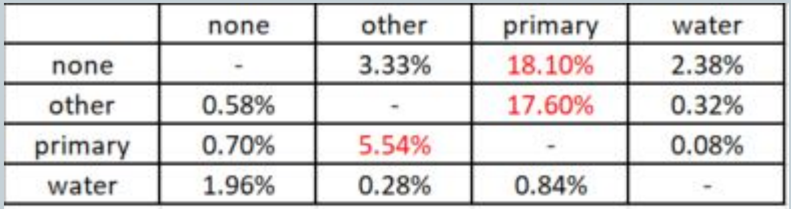


Cross-Validation	0.77 (+/- 0.00)

Confusion Matrices

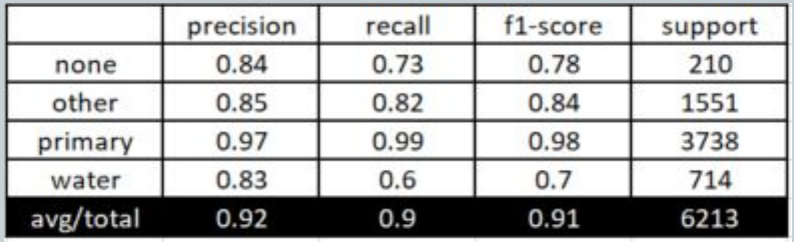


	none	other	primary	water
none	-	7	38	5
other	9	-	273	5
primary	26	207	-	3
water	14	2	6	-

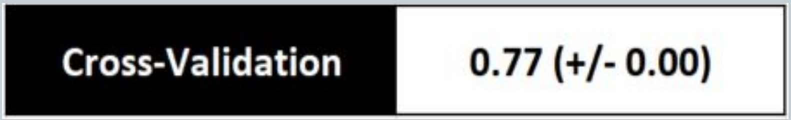


	none	other	primary	water
none	-	3.33%	18.10%	2.38%
other	0.58%	-	17.60%	0.32%
primary	0.70%	5.54%	-	0.08%
water	1.96%	0.28%	0.84%	-

Classification Report

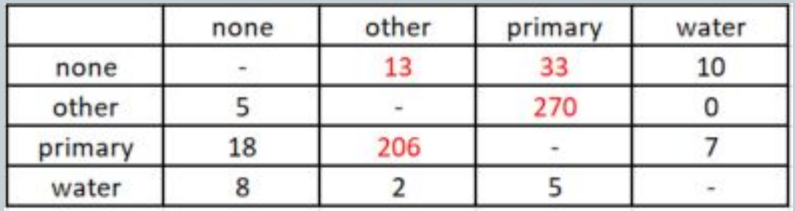


	precision	recall	f1-score	support
none	0.84	0.73	0.78	210
other	0.85	0.82	0.84	1551
primary	0.97	0.99	0.98	3738
water	0.83	0.6	0.7	714
avg/total	0.92	0.9	0.91	6213

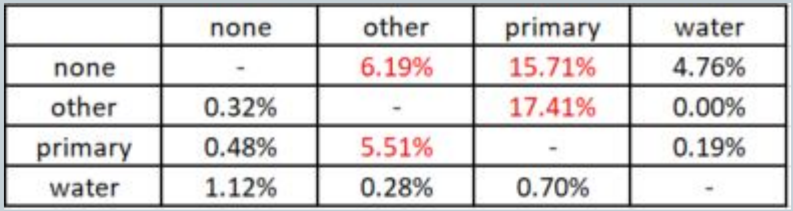


Cross-Validation	0.77 (+/- 0.00)

Confusion Matrices



	none	other	primary	water
none	-	13	33	10
other	5	-	270	0
primary	18	206	-	7
water	8	2	5	-



	none	other	primary	water
none	-	6.19%	15.71%	4.76%
other	0.32%	-	17.41%	0.00%
primary	0.48%	5.51%	-	0.19%
water	1.12%	0.28%	0.70%	-

Conclusion

- Our Weather Models were reasonably accurate except for haze prediction
- Our land model worked well for primary and human features, but recall for water is something we are still trying to improve
- Still - we were accurately predicting the primary model >97% of the time on both precision and recall
- This means our model is very good at doing what it is intended to - identifying when forest is there or not

Questions

?????