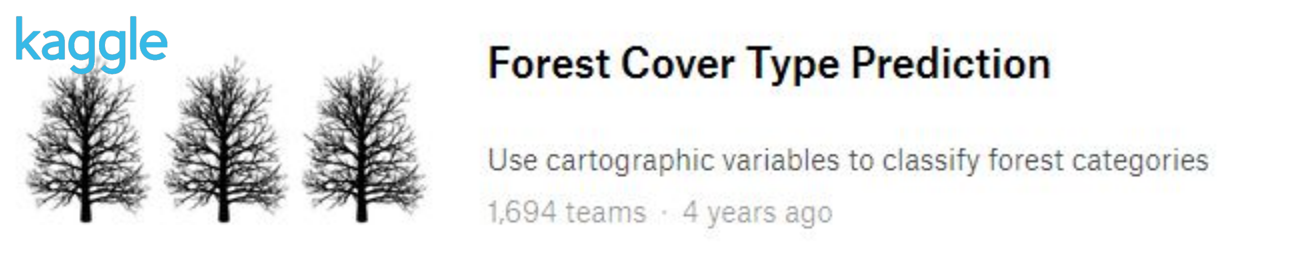
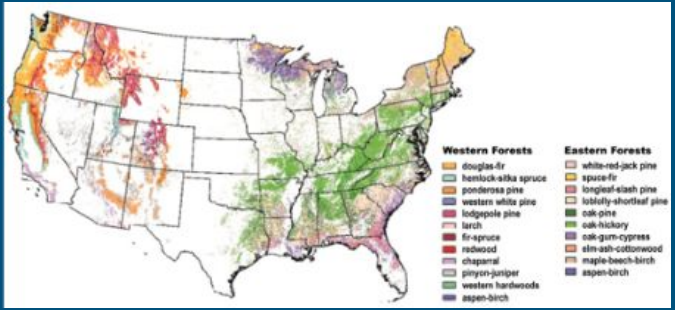


Forest Cover Type Prediction

Use cartographic variables to classify forest categories

1,694 teams · 4 years ago

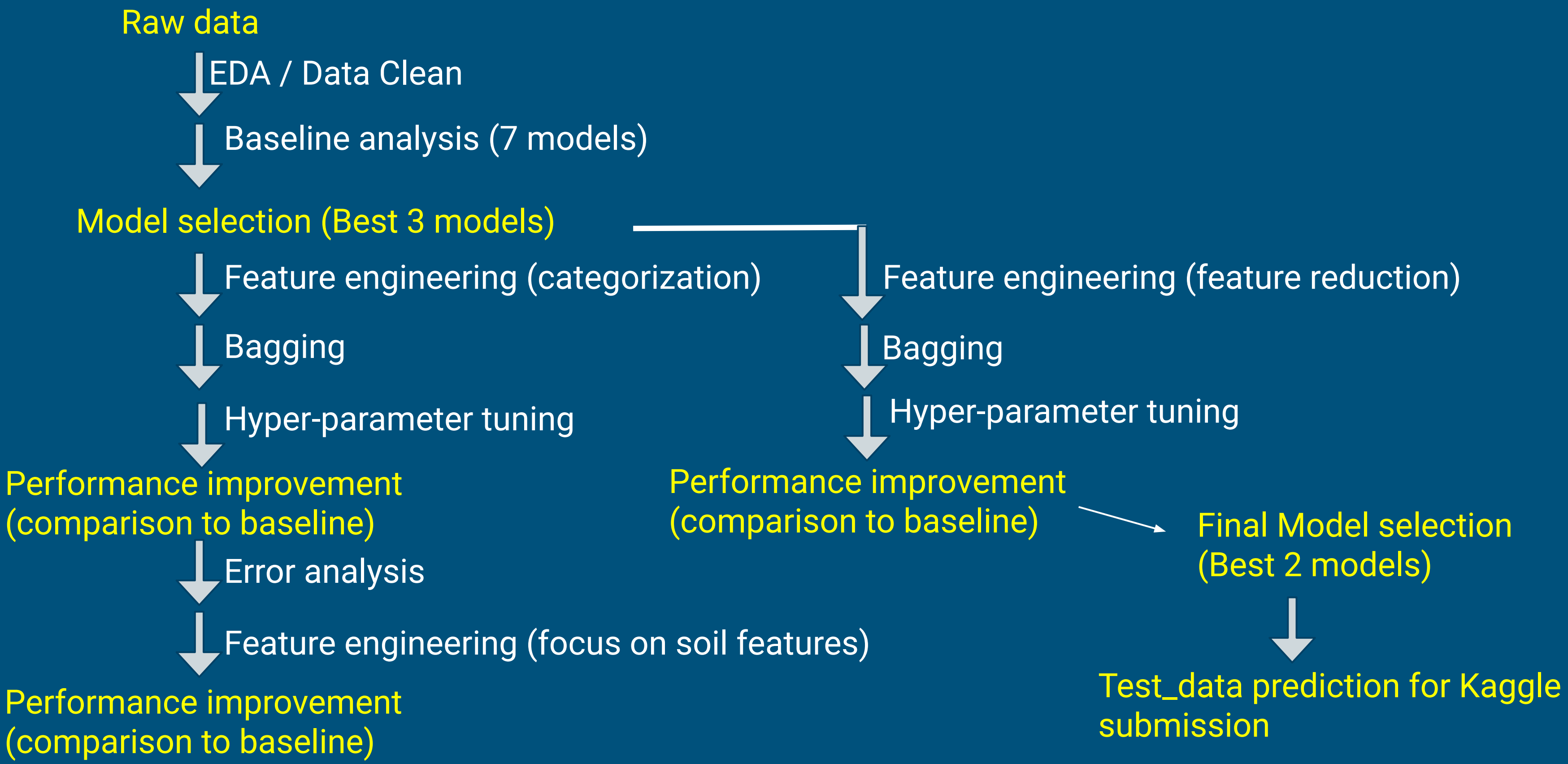
Chloe Wu, Erik Hou, Adam Sohn, Curtis Lin



Motivation - Aid conservation efforts

- Forests are 740 million acres of United States ~ one third of total land.
- Forests are homes for a myriad of species.
- Forests can be strengthened by sustainable usage and weakened by wildfire and development.
- Developing an **machine learning-based forest classification tool** provides conservationists a tool to leverage minimal data to better characterize the environment.

Data analysis plan



Raw data

EDA / Data Clean

Baseline analysis (7 models)

Model selection (Best 3 models)

Feature engineering (categorization)

Bagging

Hyper-parameter tuning

Performance improvement
(comparison to baseline)

Error analysis

Feature engineering (focus on soil features)

Performance improvement
(comparison to baseline)

Feature engineering (feature reduction)

Bagging

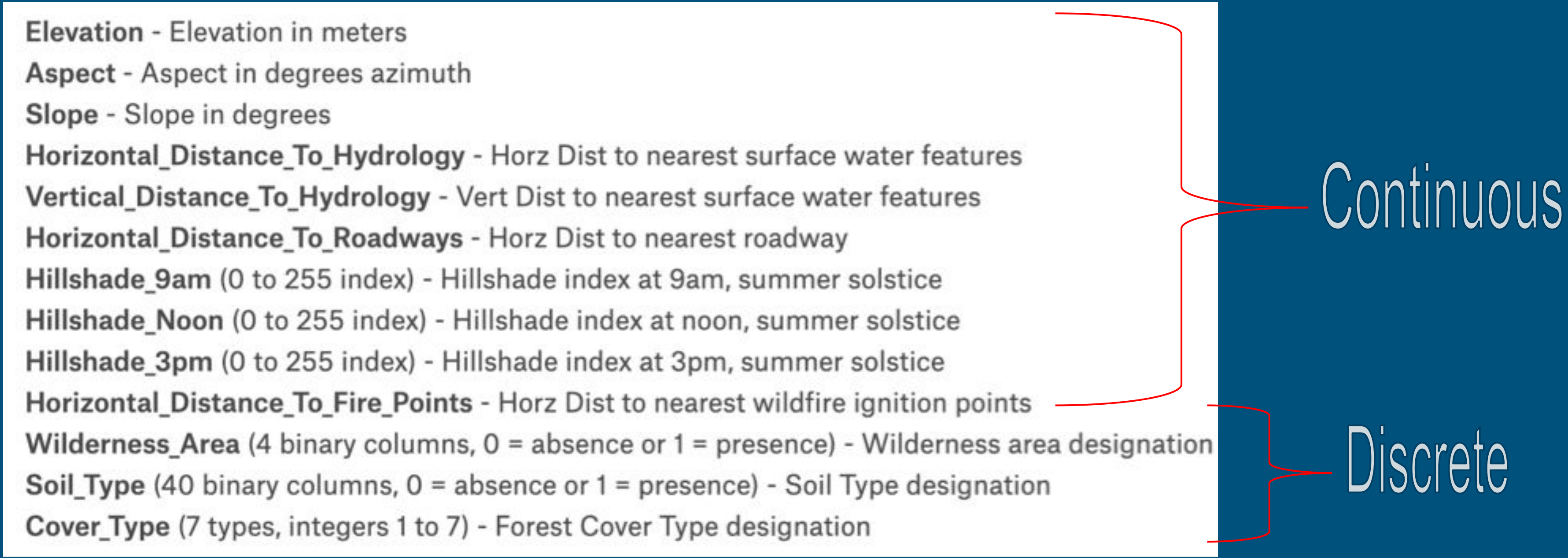
Hyper-parameter tuning

Performance improvement
(comparison to baseline)

Final Model selection
(Best 2 models)

Test_data prediction for Kaggle submission

EDA



Elevation - Elevation in meters

Aspect - Aspect in degrees azimuth

Slope - Slope in degrees

Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features

Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features

Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway

Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice

Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice

Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice

Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points

Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation

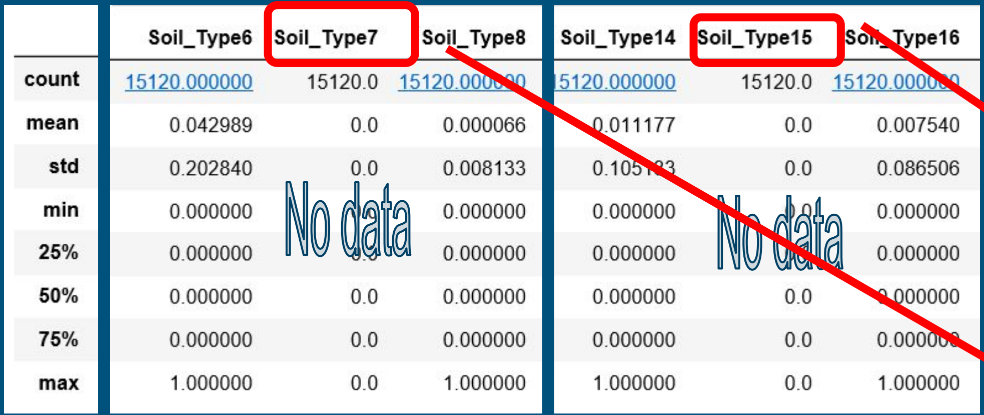
Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation

Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation

Continuous

Discrete

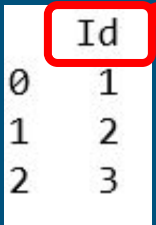
Data Clean



	Soil_Type6	Soil_Type7	Soil_Type8	Soil_Type14	Soil_Type15	Soil_Type16
count	15120.000000	15120.0	15120.000000	15120.000000	15120.0	15120.000000
mean	0.042989	0.0	0.000066	0.011177	0.0	0.007540
std	0.202840	0.0	0.008133	0.105133	0.0	0.086506
min	0.000000	0.0	0.000000	0.000000	0.0	0.000000
25%	0.000000	0.0	0.000000	0.000000	0.0	0.000000
50%	0.000000	0.0	0.000000	0.000000	0.0	0.000000
75%	0.000000	0.0	0.000000	0.000000	0.0	0.000000
max	1.000000	0.0	1.000000	1.000000	0.0	1.000000

No data

No data



	Id
0	1
1	2
2	3

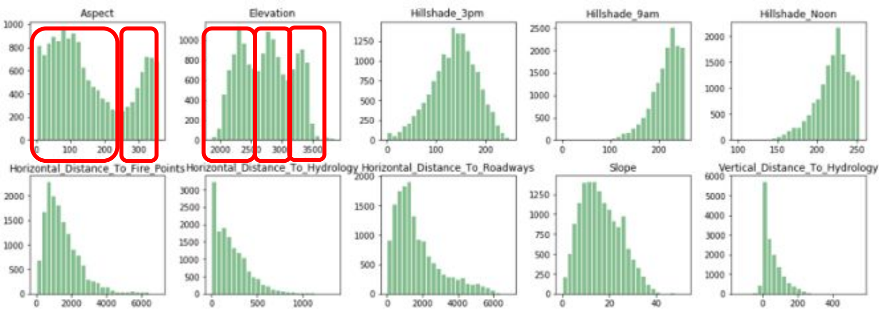
Sequential

Baseline Analysis

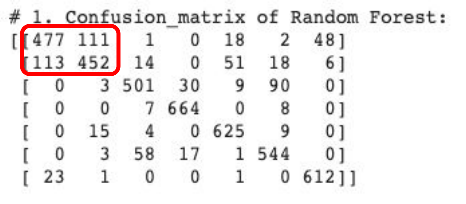
Classifier	Accuracy
Support Vector Machine (SVM) Un-normalized	0.8488
Random Forest	0.8208
Decision Tree	0.7923
K-Nearest Neighbors	0.7919
Logistic Regression	0.6759
Support Vector Model (SVM) Normalized	0.6232
Gaussian Naive Bayes	0.5983

Minimal hyperparameter tuning & no Feature Engineering

Feature Engineering - Categorization

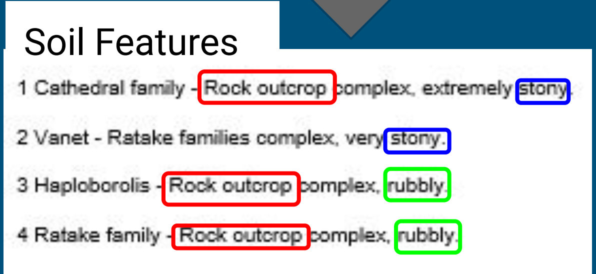


Error Analysis



```
# 1. Confusion_matrix of Random Forest:
[[477 111   1   0  18   2  48]
 [113 452  14   0  51  18   6]
 [  0   3 501  30   9  90   0]
 [  0   0   7 664   0   8   0]
 [  0  15   4   0 625   9   0]
 [  0   3  58  17   1 544   0]
 [ 23   1   0   0   1   0 612]]
```

Adding Features



Soil Features

1 Cathedral family - Rock outcrop complex, extremely stony.

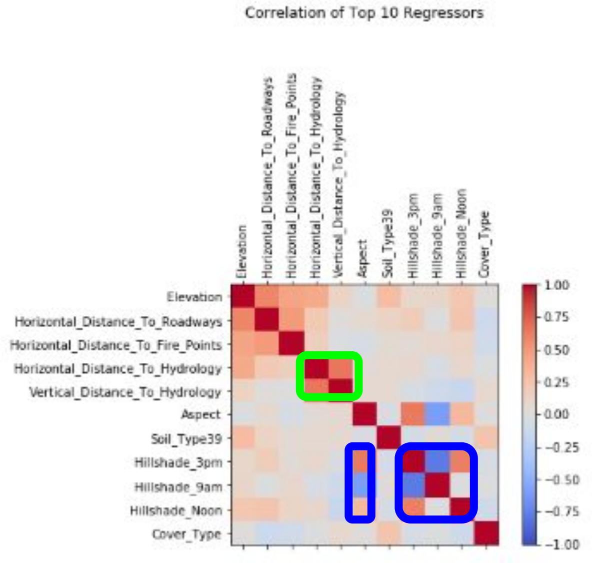
2 Vanet - Ratake families complex, very stony.

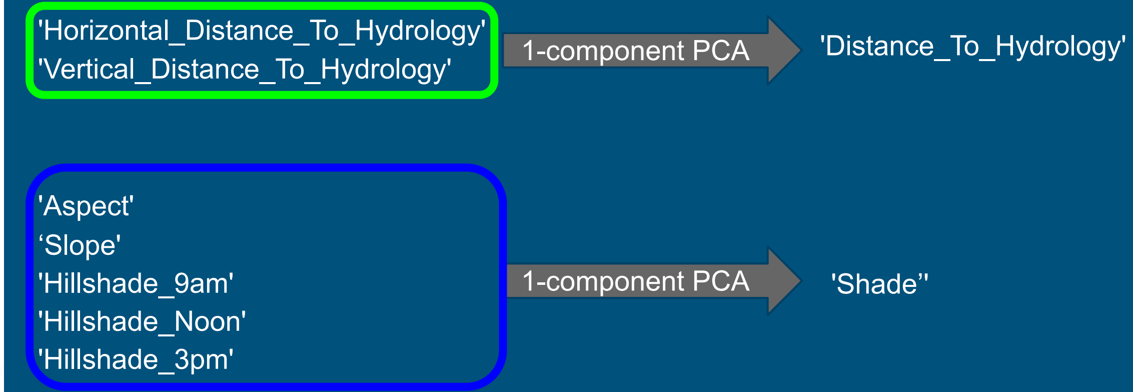
3 Haploborolis - Rock outcrop complex, rubbly.

4 Ratake family - Rock outcrop complex, rubbly.

Improvement (Max delta)
Random Forest: 0.8208 (baseline) -> 0.8302 (no bagging) -> N/A (w/ bagging) -> 0.8543 (w/ hyperparameter tuning)= **+ 3.35%**
SVM: 0.8488 (baseline) -> 0.8492(no bagging) -> 0.8415(w/ bagging) -> 0.8611 (w/ hyperparameter tuning) = **+ 1.23%**
Decision Tree: 0.7923 (baseline) -> 0.7897(no bagging) -> 0.8646 (w/ Adaboost) -> **0.8560** (w/ hyperparameter tuning) = **+ 7.23%**

Feature Engineering - Feature Reduction





'Horizontal_Distance_To_Hydrology'
'Vertical_Distance_To_Hydrology'

1-component PCA

'Distance_To_Hydrology'

'Aspect'
‘Slope'
'Hillshade_9am'
'Hillshade_Noon'
'Hillshade_3pm'

1-component PCA

'Shade''

<u>Improvement (Max delta)</u>

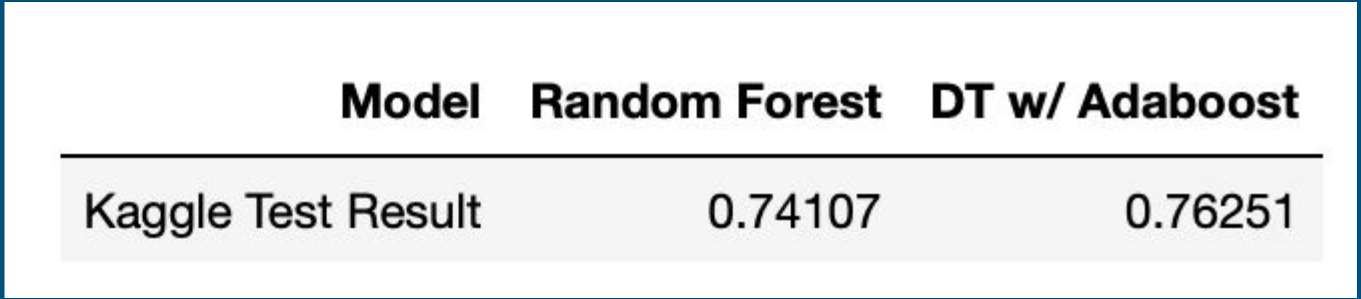
Random Forest: 0.8208 (baseline) -> 0.8781 (no bagging) -> N/A (w/ bagging) -> 0. 8796(w/ hyperparameter tuning)= **+ 5.88%**
SVM: 0.8488 (baseline) -> 0.8479 (no bagging) -> N/A(w/ bagging) -> 0.8479 (w/ hyperparameter tuning) = **- 0.09%**
Decision Tree: 0.7923 (baseline) -> N/A (no bagging) -> 0.8805(w/ Adaboost) -> **0.8814**(Adaboost w/ hyperparameter tuning) = **+ 8.91%**

Hyperparameter Tuning

Classifier	Baseline	After Tuning
Random Forest	0.8208	0.8796
Decision Tree w/ AdaBoost	0.7923 (no AdaBoost)	0.8814

Hyperparameter Modified:
Modified # of estimators, Max depth, Learning Rate

Kaggle Submission



Model	Random Forest	DT w/ Adaboost
Kaggle Test Result	0.74107	0.76251

Conclusion

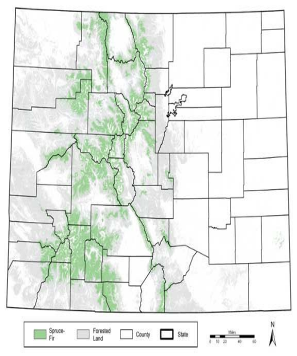
Random Forest and Decision Tree were optimal due to their abilities to navigate complex decision boundaries.

Still, classifiers had difficulty disambiguating Cover Type 1 & 2.

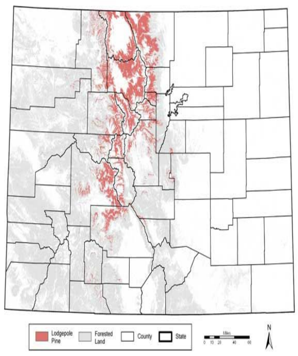
Next recommended steps are:

- Address possible over-fit
- Modify training data set
  - change train/dev %, perhaps eliminate dev set for final step.
  - ensemble learn w/ split training set
- Seek location data such as latitude and longitude

Cover Type 1



Cover Type 2


Lodgepole Pine
Forested Land
County
State
Miles
0 10 20 40 60
N