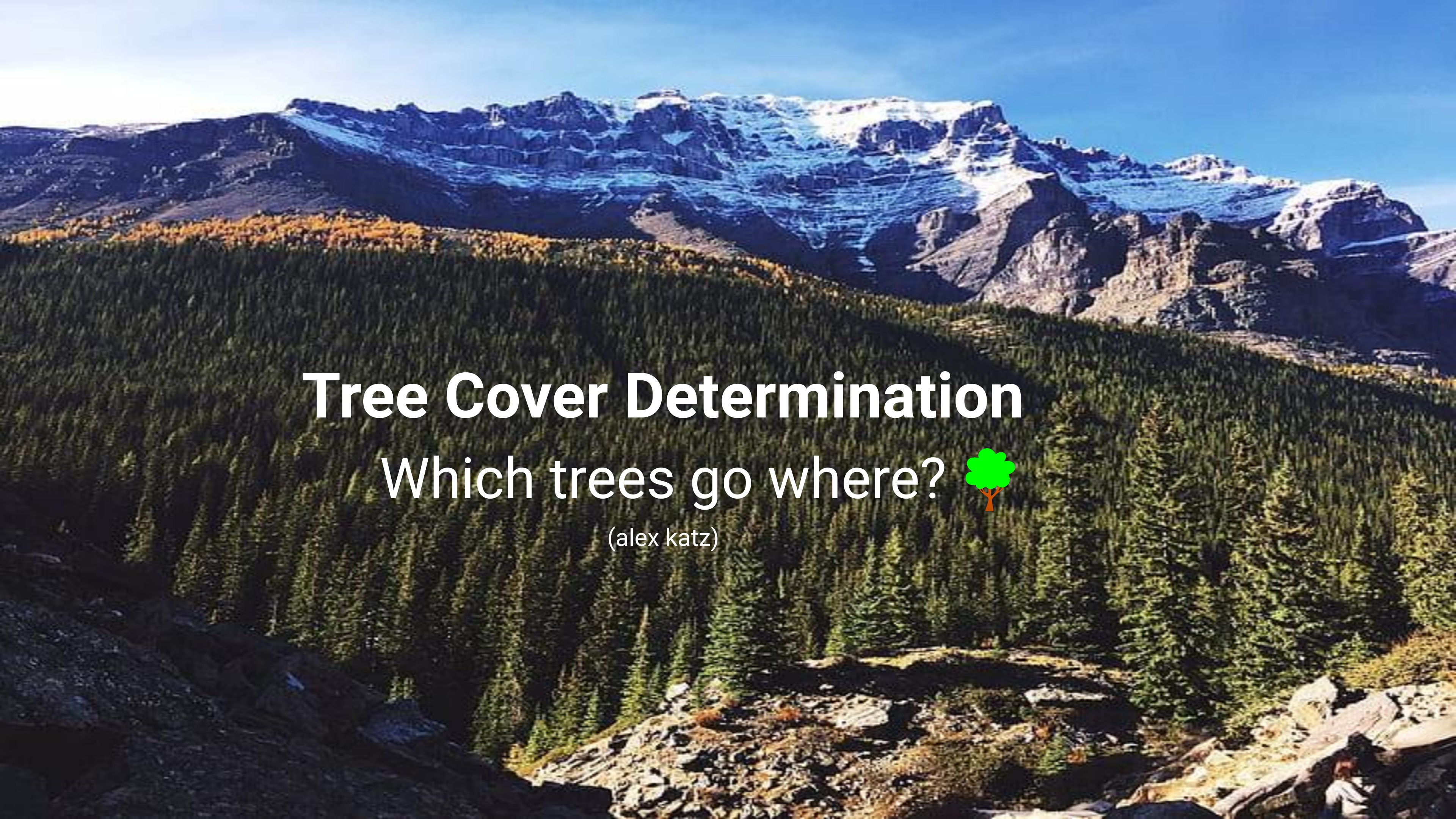


Tree Cover Determination

Which trees go where? 🌳

(alex katz)

Problem Statement

- Data from Kaggle competition (15,000 30m^2 plots of land in Roosevelt Nat. Forest)
- Theorize use case scenario:
  - Land heavily logged in 1990s
  - Recent forest fires have cleared land of vegetation
  - Colorado wants to repopulate tree cover

Features and Organization

Target Variable

Cover Type

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

Quantitative Features

- Elevation
- Aspect
- Slope
- Horizontal Distance To Hydrology
- Vertical Distance To Hydrology
- Horizontal Distance To Roadways
- Horizontal_Distance_To_Fire_Points
- Hillshade_9am
- Hillshade_Noon
- Hillshade_3pm

Qualitative Features

Wilderness Area (4)

1 - Rawah Wilderness Area

2 - Neota Wilderness Area

3 - Comanche Peak Wilderness Area

4 - Cache la Poudre Wilderness Area

Soil Type (40)

Methodology

Feature Engineering and EDA

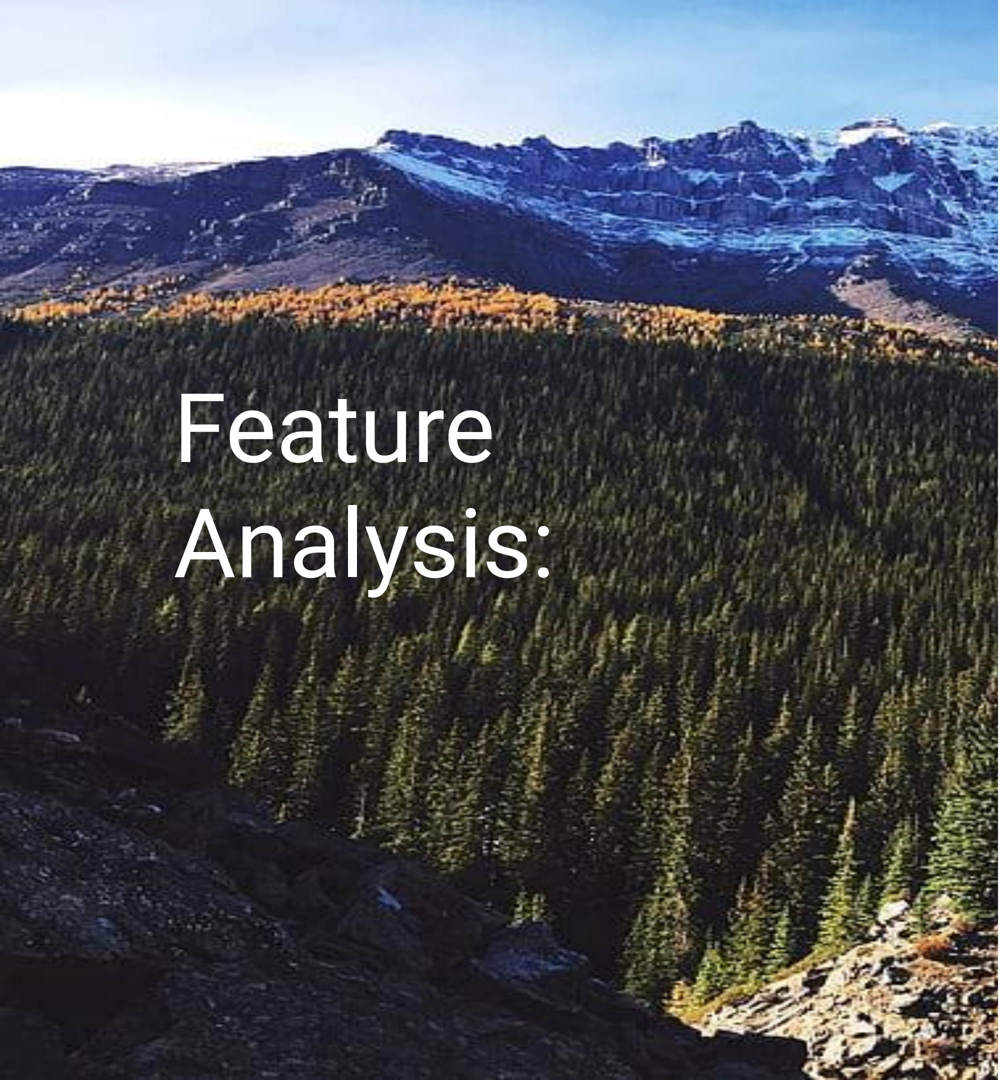
- Remove some collinear features
- Use log of features if highly skewed
- Euclidean distance
- Additional feature relationships, etc.

Model Development and Selection

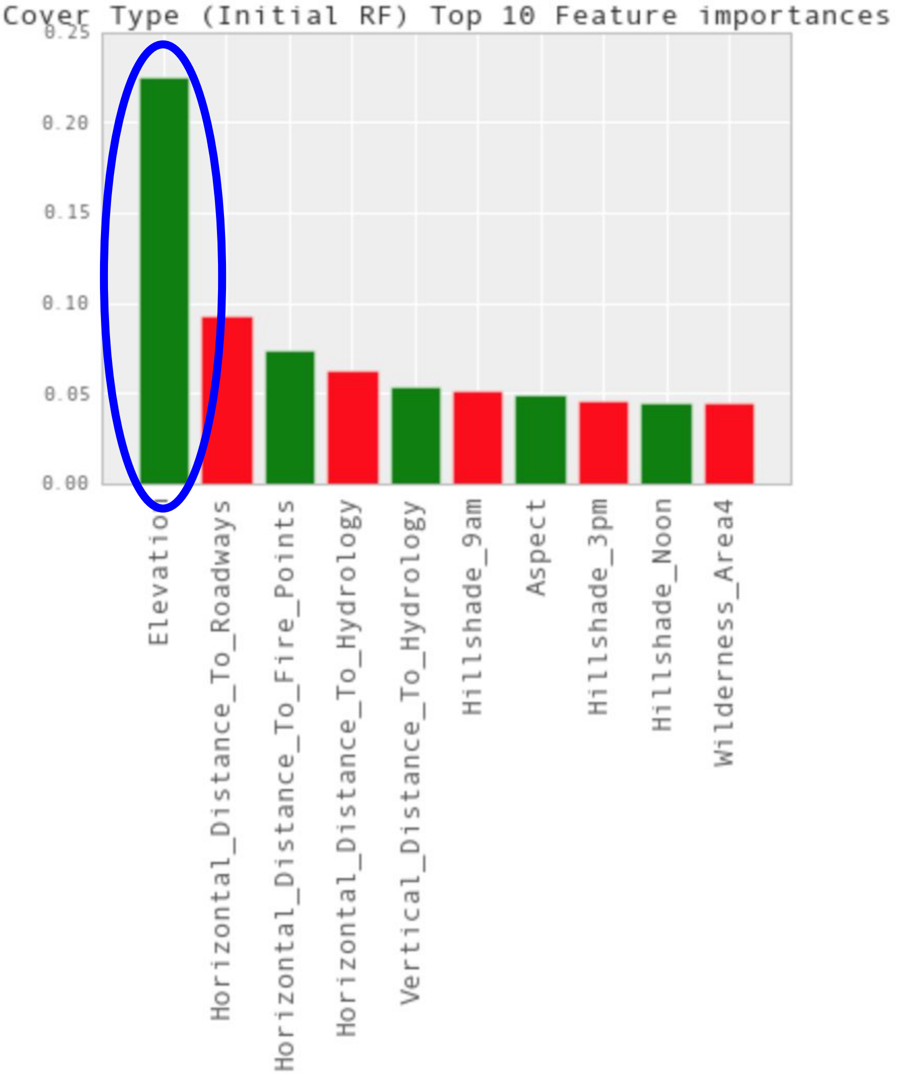
- Logistic Regression
- SVC
- Decision Tree
- Random Forests
- XGBoost
- KNN

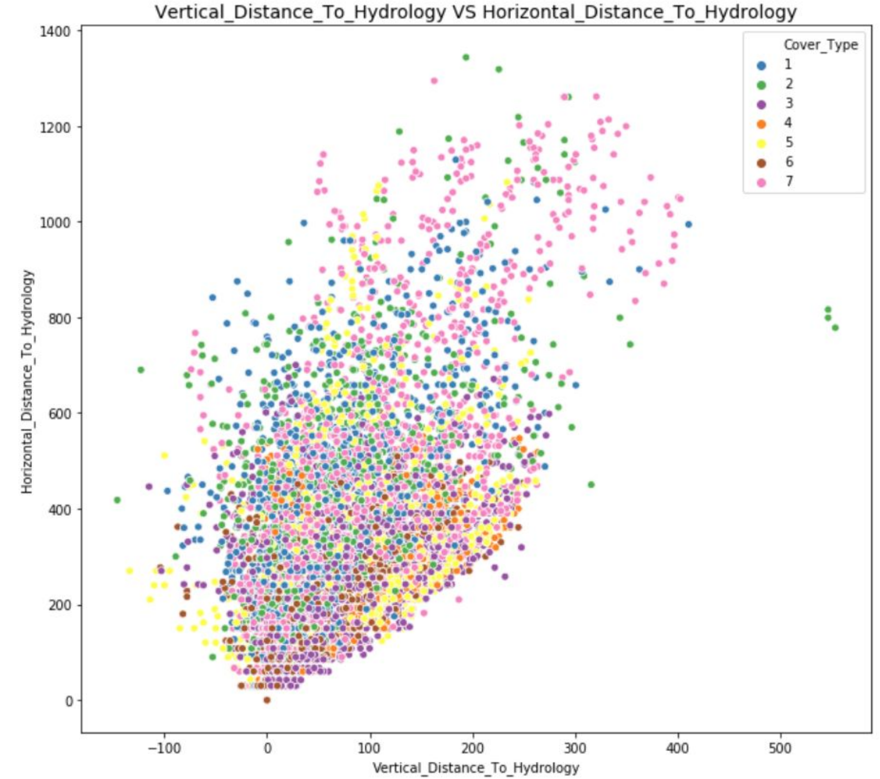
Model Tuning and Analysis

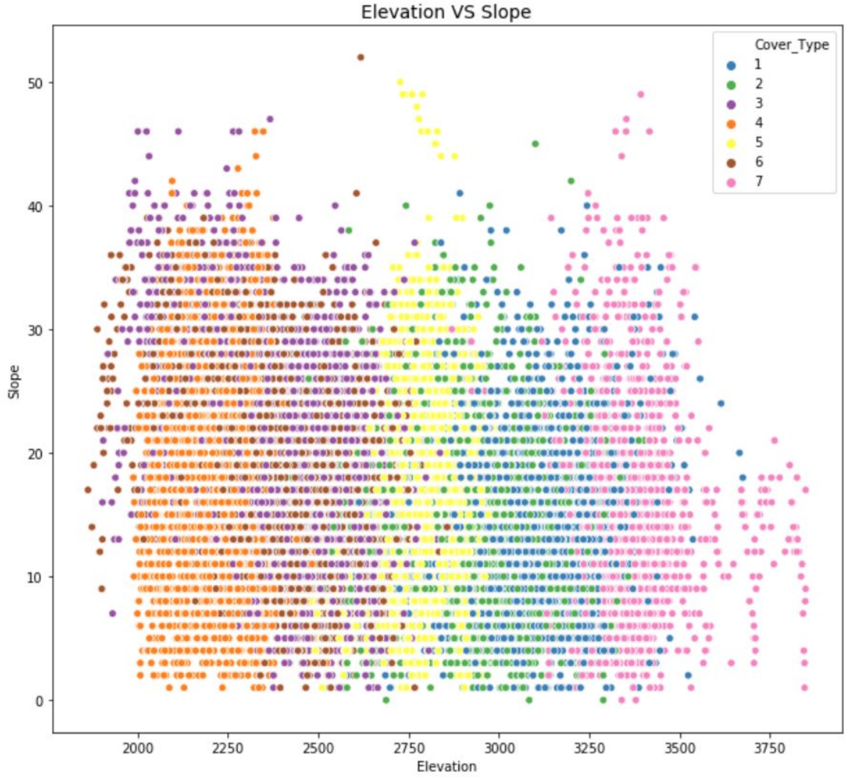
- Maximize Recall
- RandomSearch and GridSearch to tune parameters for models

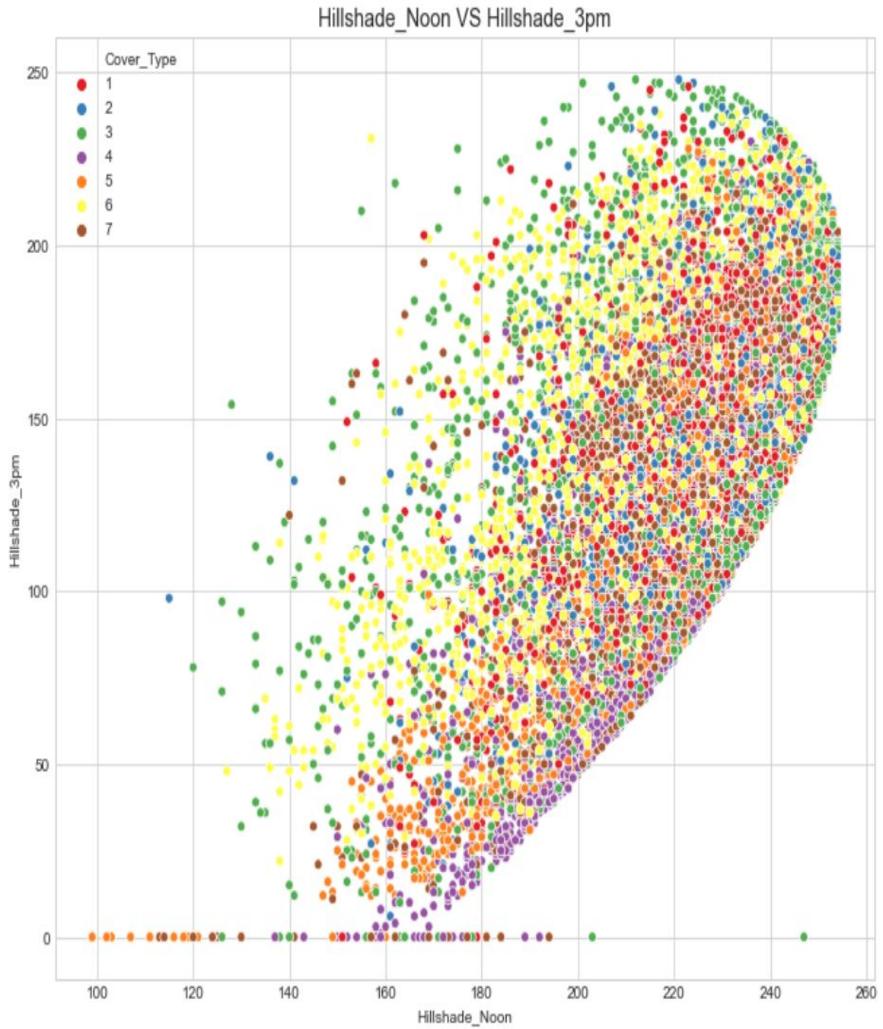


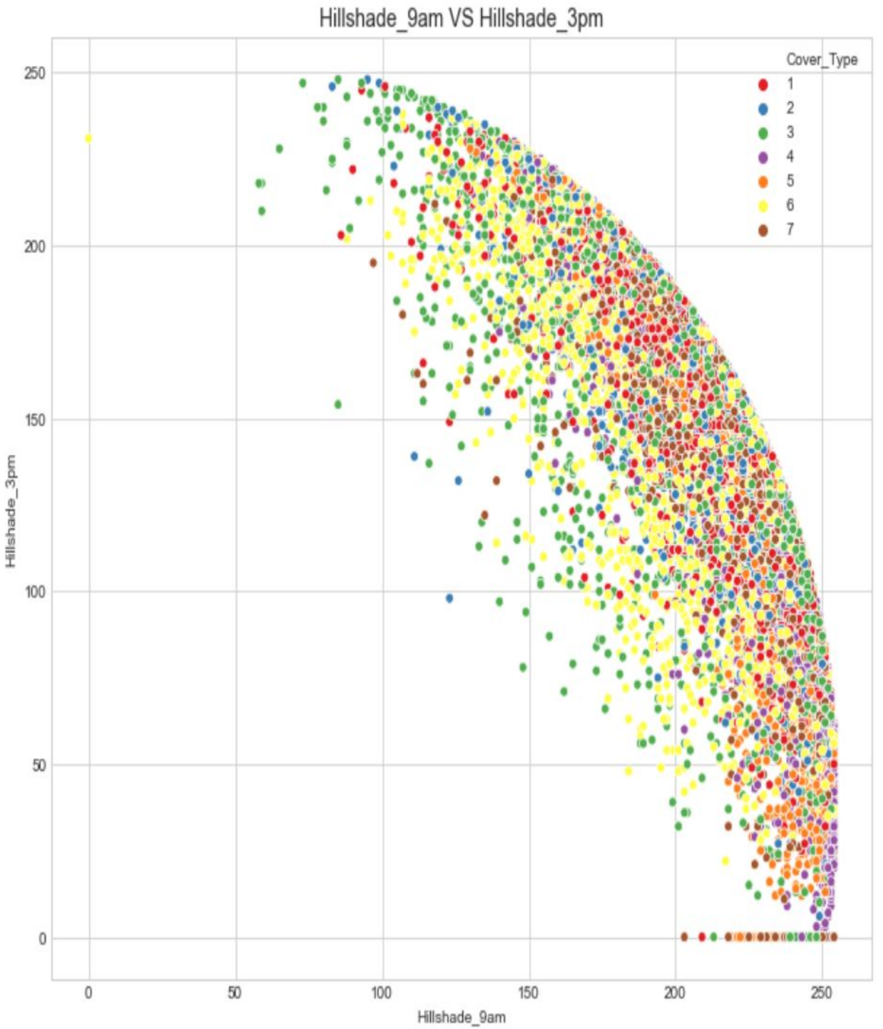
Feature Analysis:



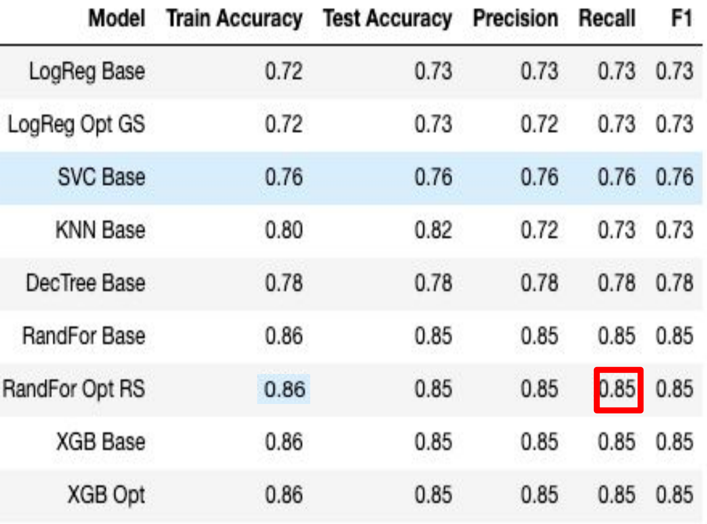

Vertical_Distance_To_Hydrology VS Horizontal_Distance_To_Hydrology
Horizontal_Distance_To_Hydrology
1400
1200
1000
800
600
400
200
0
−100
0
100
200
300
400
500
Vertical_Distance_To_Hydrology
Cover_Type
● 1
● 2
● 3
● 4
● 5
● 6
● 7


Elevation VS Slope
Slope
50
40
30
20
10
0
2000
2250
2500
2750
3000
3250
3500
3750
Elevation
Cover_Type
● 1
● 2
● 3
● 4
● 5
● 6
● 7

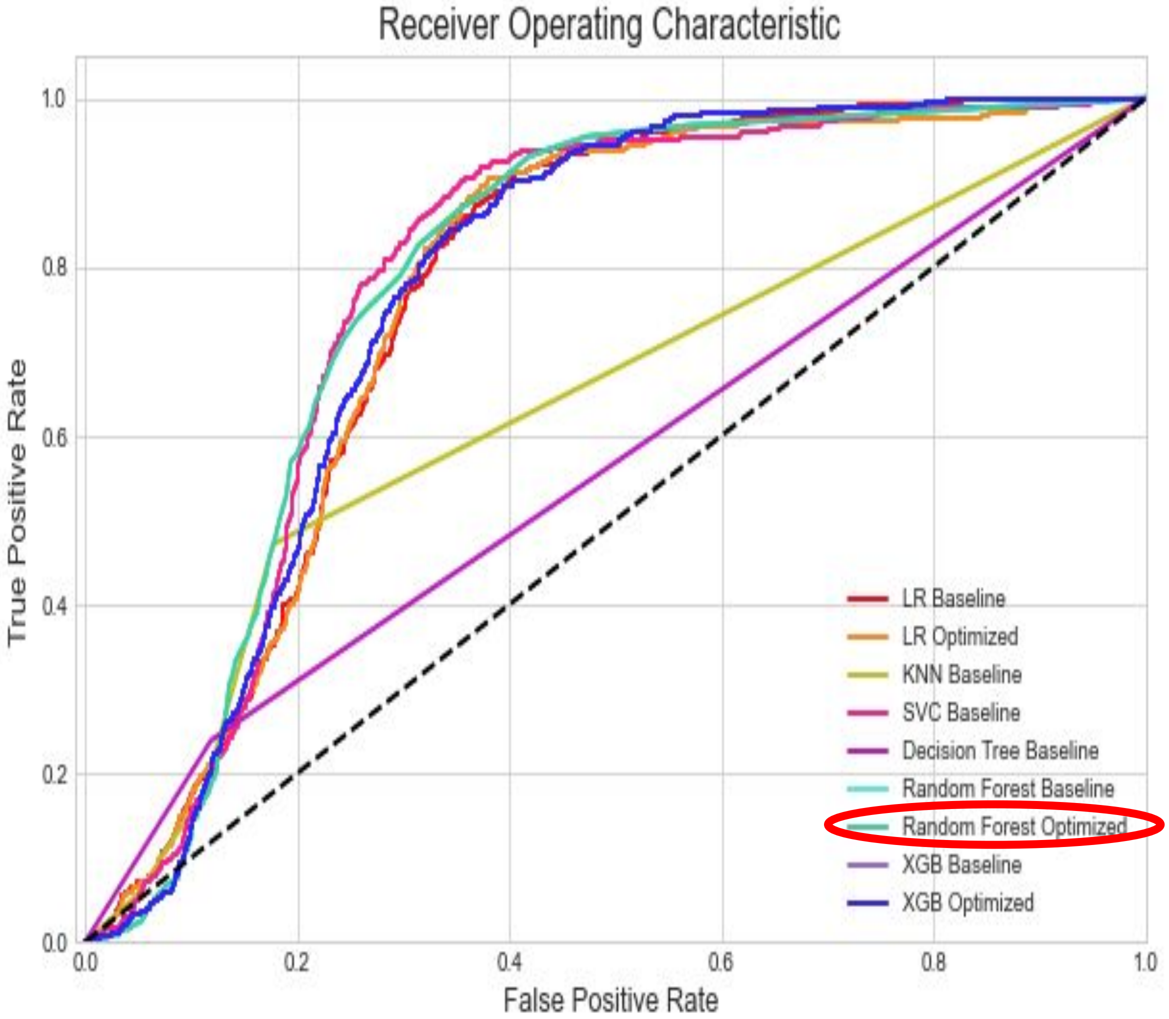

Hillshade_Noon VS Hillshade_3pm
Cover_Type
1
2
3
4
5
6
7
Hillshade_3pm
0
50
100
150
200
250
Hillshade_Noon
100
120
140
160
180
200
220
240
260

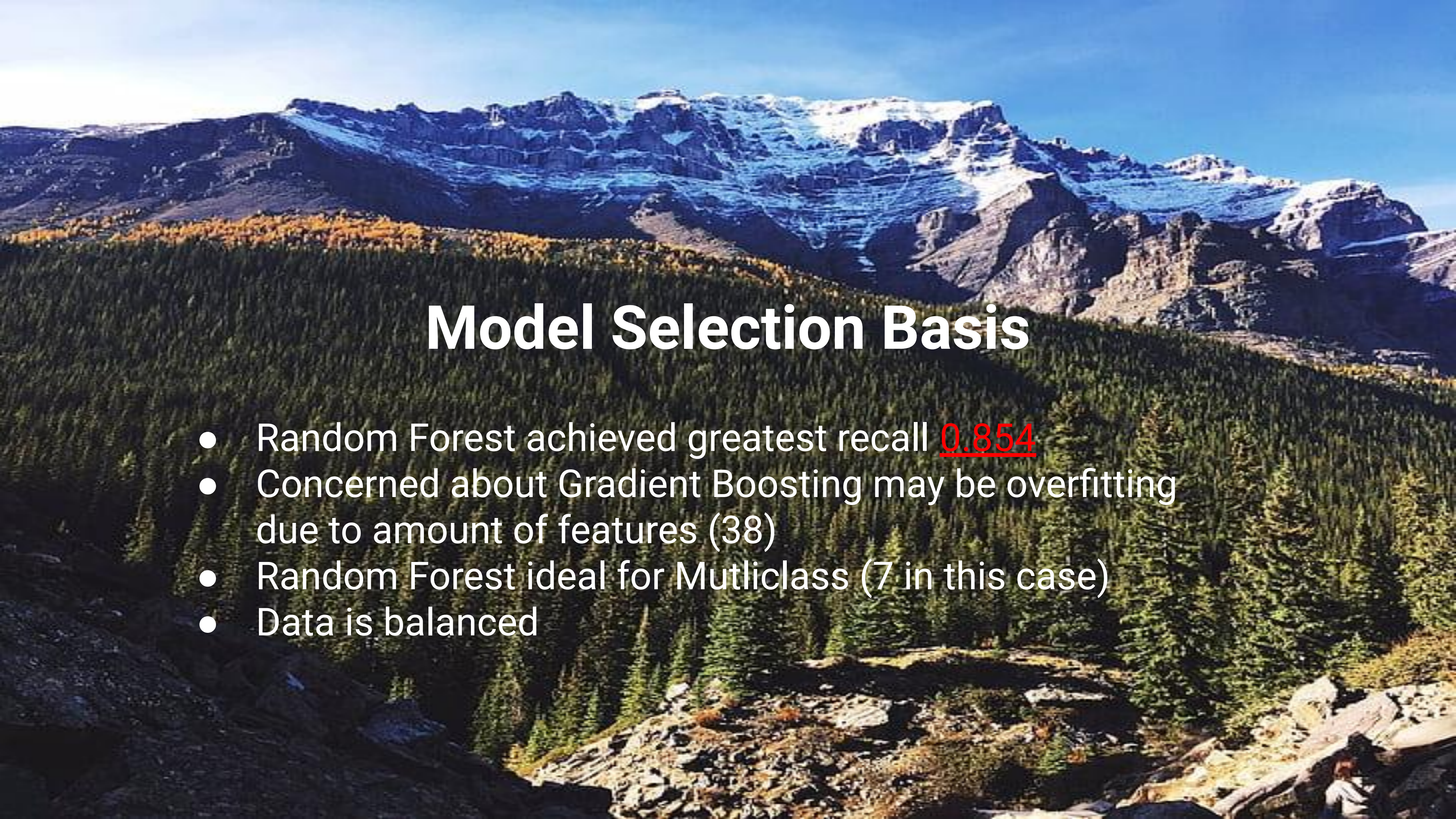

Hillshade_9am VS Hillshade_3pm
Cover_Type
1
2
3
4
5
6
7
Hillshade_3pm
0
50
100
150
200
250
Hillshade_9am
0
50
100
150
200
250

Models Comparison



Model	Train Accuracy	Test Accuracy	Precision	Recall	F1
LogReg Base	0.72	0.73	0.73	0.73	0.73
LogReg Opt GS	0.72	0.73	0.72	0.73	0.73
SVC Base	0.76	0.76	0.76	0.76	0.76
KNN Base	0.80	0.82	0.72	0.73	0.73
DecTree Base	0.78	0.78	0.78	0.78	0.78
RandFor Base	0.86	0.85	0.85	0.85	0.85
RandFor Opt RS	0.86	0.85	0.85	0.85	0.85
XGB Base	0.86	0.85	0.85	0.85	0.85
XGB Opt	0.86	0.85	0.85	0.85	0.85





Model Selection Basis

- Random Forest achieved greatest recall 0.854
- Concerned about Gradient Boosting may be overfitting due to amount of features (38)
- Random Forest ideal for Mutliclass (7 in this case)
- Data is balanced

Optimized Scaled RandomForest:

Precision: 0.85

Recall: 0.85

Parameters: n_estimators=12,
criterion='entropy',
max_features=6,
max_depth=None,
min_samples_split=100,
min_samples_leaf=20,
bootstrap=False,
random_state=1)

Good Performers:

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

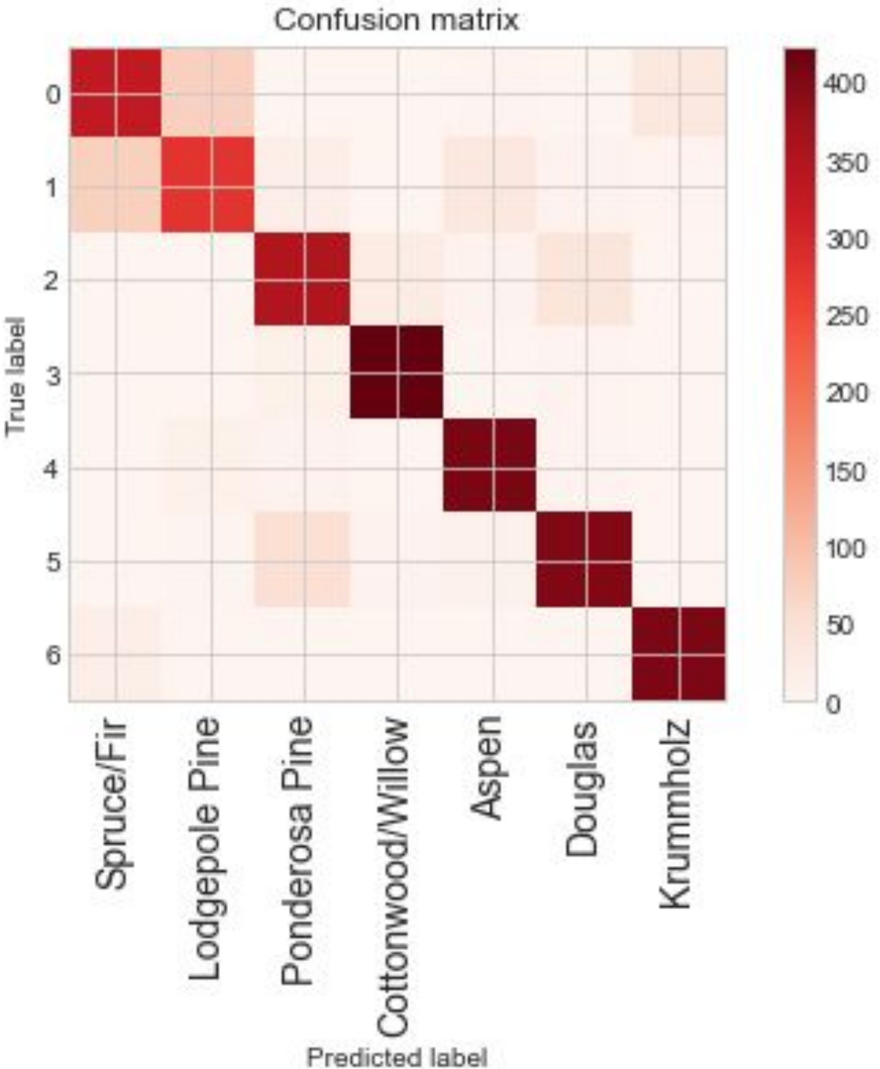
7 - Krummholz

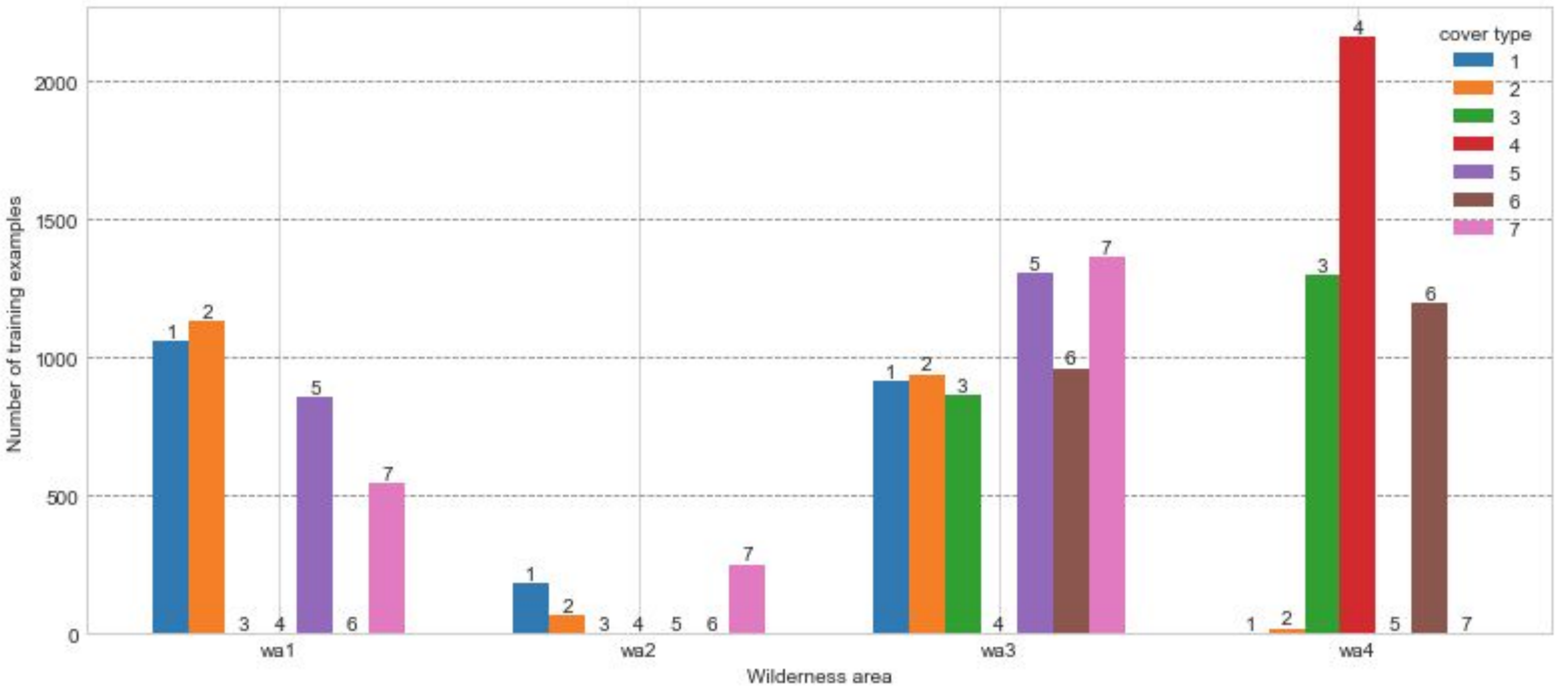
Poor Performers:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine





Some Insights:

- We should never classify Tree Cover in area 4 as 1 or 2 (12)
- We should never classify Tree Cover in area 1 as 3, 4 or 6 (15)
- Hesitant to make the same assumption for area 2

Conclusions and Takeaways

- Current Random Forest Model is decent, but could be improved (feature engineering)
- Account for class imbalance which would exist in real world application
- Consider breaking down by Wilderness Area/Region in larger data set

Other Use Cases Examples:

- Identifying where invasive species are located
- Logging (hopefully not)
- Reverse the logic-tree cover helps identify some features