Predicting Forest Cover Types within the Roosevelt National Forest

Lauren Aronson, ASA

Content

- Predicting Forest Cover Types within the Roosevelt National Forest
  - Background & Methodology
  - Findings & Recommendations
  - Future Work & Next Steps

- Questions

- Appendix

Background & Methodology

Purpose

- ➢ Machine Learning project with Flatiron School
  - Select a dataset for classification modeling

- ➢ Predicting forest cover types within Roosevelt National Forest
  - Environmental sustainability
  - Understand the ecosystem
  - Conservation efforts
  - Colorado centered

Data

DATABASE DETAILS	
Data Sourced:	From Kaggle as part of the UCI Machine Learning Repository *Original Owners: Remote Sensing and GIS Program, Department of Forest Sciences, College of Natural Resources, Colorado State University*
Data Determined By:	US Forest Service (USFS) Region 2 Resource Information System (RIS)
Date Donated:	1998-08-01
Dataset Characteristics:	Multivariate
Attribute Characteristics:	Categorical, Integer
Associated Tasks:	Classification
Number of Instances:	581,012
Number of Attributes:	54
Missing Values?	No

Data (continued)

- **Study Area**
  - Roosevelt National Forest of Northern Colorado
  - 4 wilderness areas : Rawah; Neota; Comanche Peak; Cache la Poudre
  - Each instance represents 30m x 30m patch
- **Data Fields:**
  - Elevation - Elevation in meters
  - Aspect - Aspect in degrees azimuth
  - Slope - Slope in degrees
  - Horizontal_Distance_To_Hydrology - Horizontal Distance to nearest surface water features
  - Vertical_Distance_To_Hydrology - Vertical Distance to nearest surface water features
  - Horizontal_Distance_To_Roadways - Horizontal Distance 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 - Horizontal Distance 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
    - Spruce/Fir; Lodgepole Pine; Ponderosa Pine; Cottonwood/Willow; Aspen; Douglas-fir; Krummholz

* A summary of each forest cover type designation, wilderness area designation, and soil type designation can be found in the Appendix

Architecture

ARCHITECTURE DETAILS	
Server:	Google Colaboratory *RAM: 13GB* *Storage: 38GB* *2-core xeon 2.2GHz*
Database:	Flat File (CSV)
Programming Language:	Python 3
Machine Learning Libraries:	scikit-learn, XGBoost

Methodology & Takeaway Analysis

Methodology

- Data collected from Kaggle
- Data explored for cleaning
  - No missing data or inaccurate data records
- Data explored for analysis
- Feature Engineering
  - Correlated features removed
  - Data resampled: Smaller soil types removed (outliers) & helped normalize the data (Elevation)
  - Data standardized for modeling
- Modeling
  - K-Nearest Neighbors
  - Random Forest
  - XGBoost

Takeaway Analysis

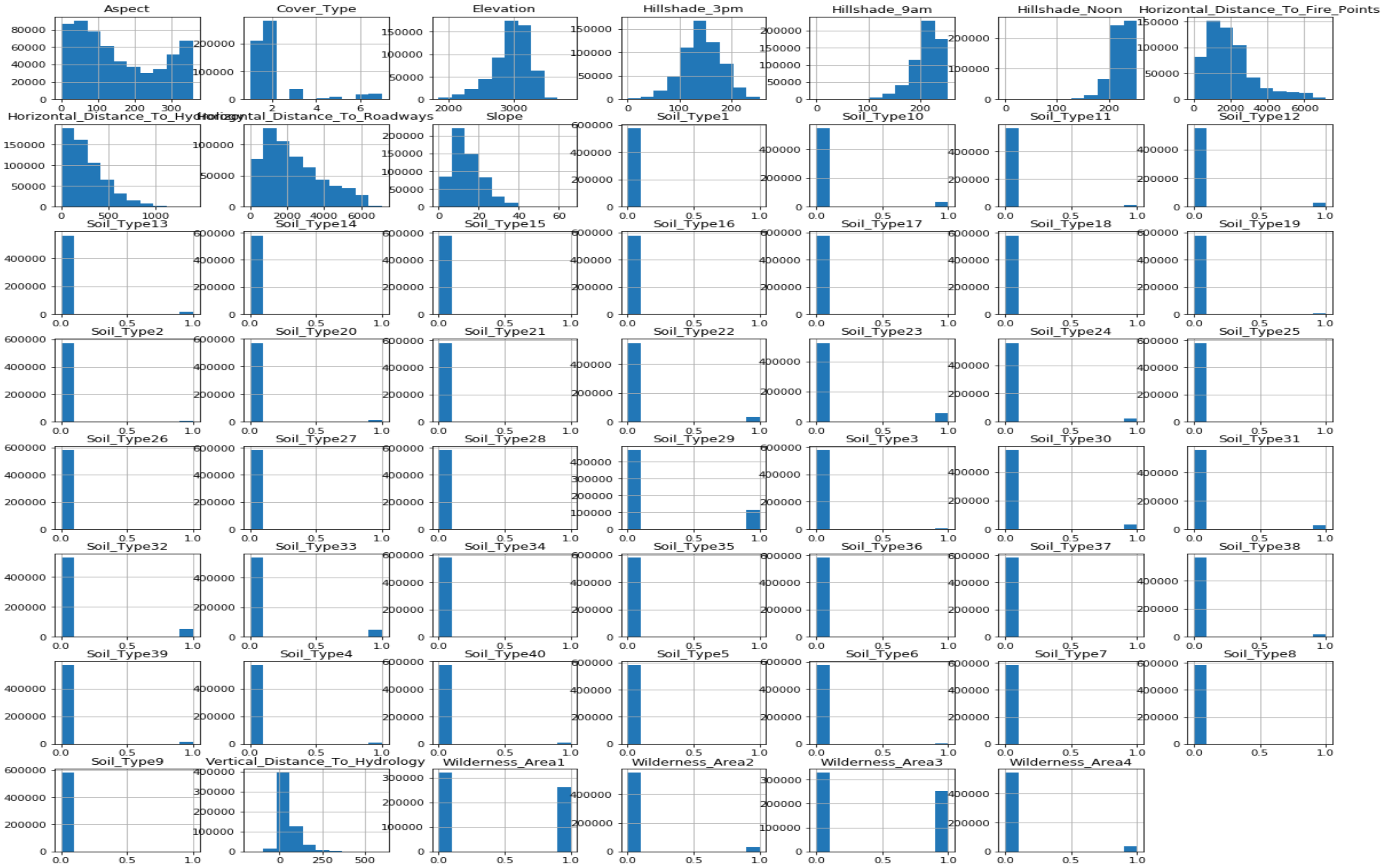
- K-Nearest Neighbors
  - Accuracy = 88%
  - Model does not provide prediction for feature importance
- Random Forest
  - Accuracy = 91%
  - Most important feature = Elevation (heavy dependence)
- XGBoost
  - Accuracy = 92%
  - Most important feature = Soil_Type12

Findings & Recommendations

Exploration

Distributions

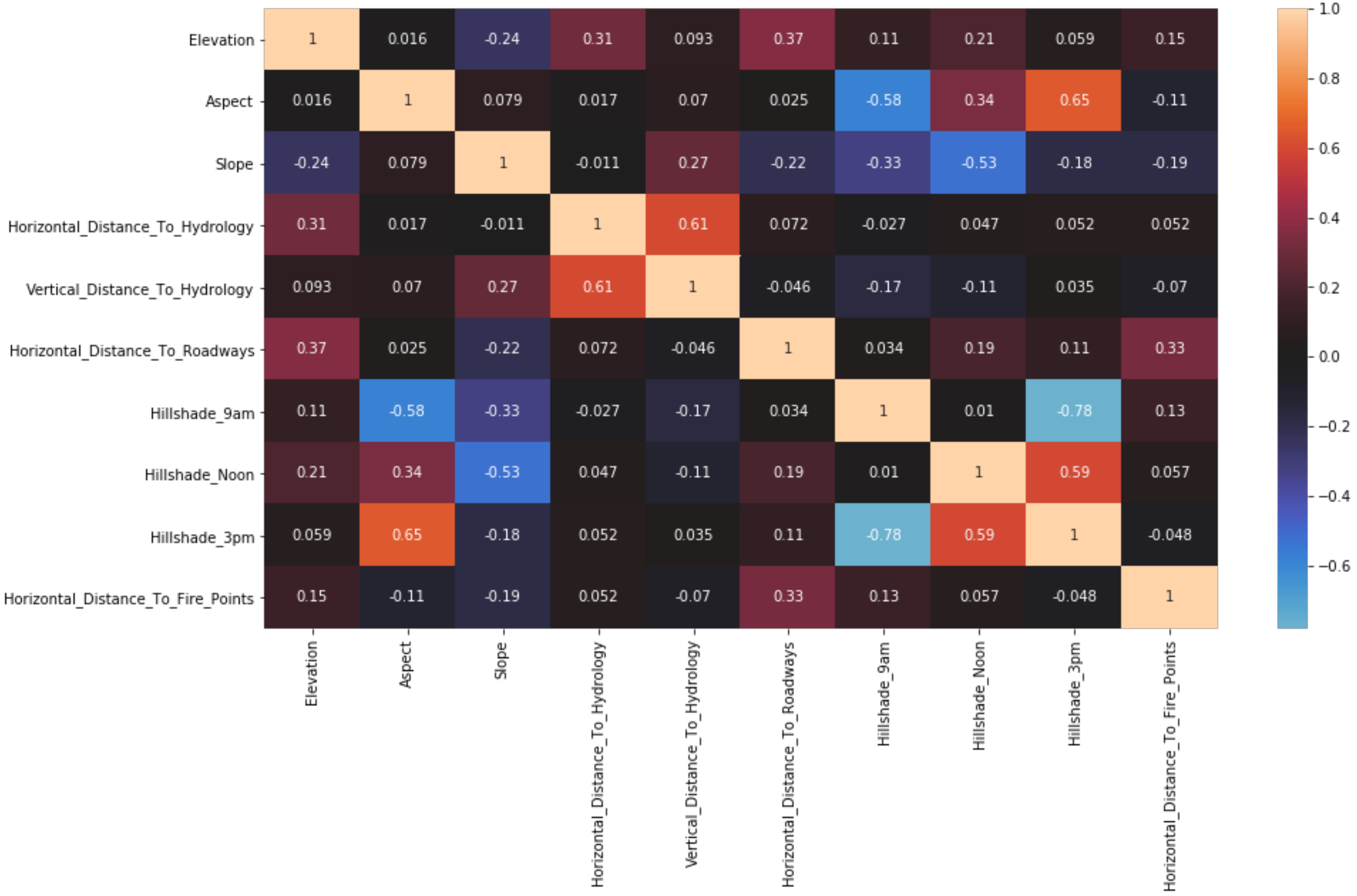
Distribution of Features



Exploration

Correlation

Correlated Features

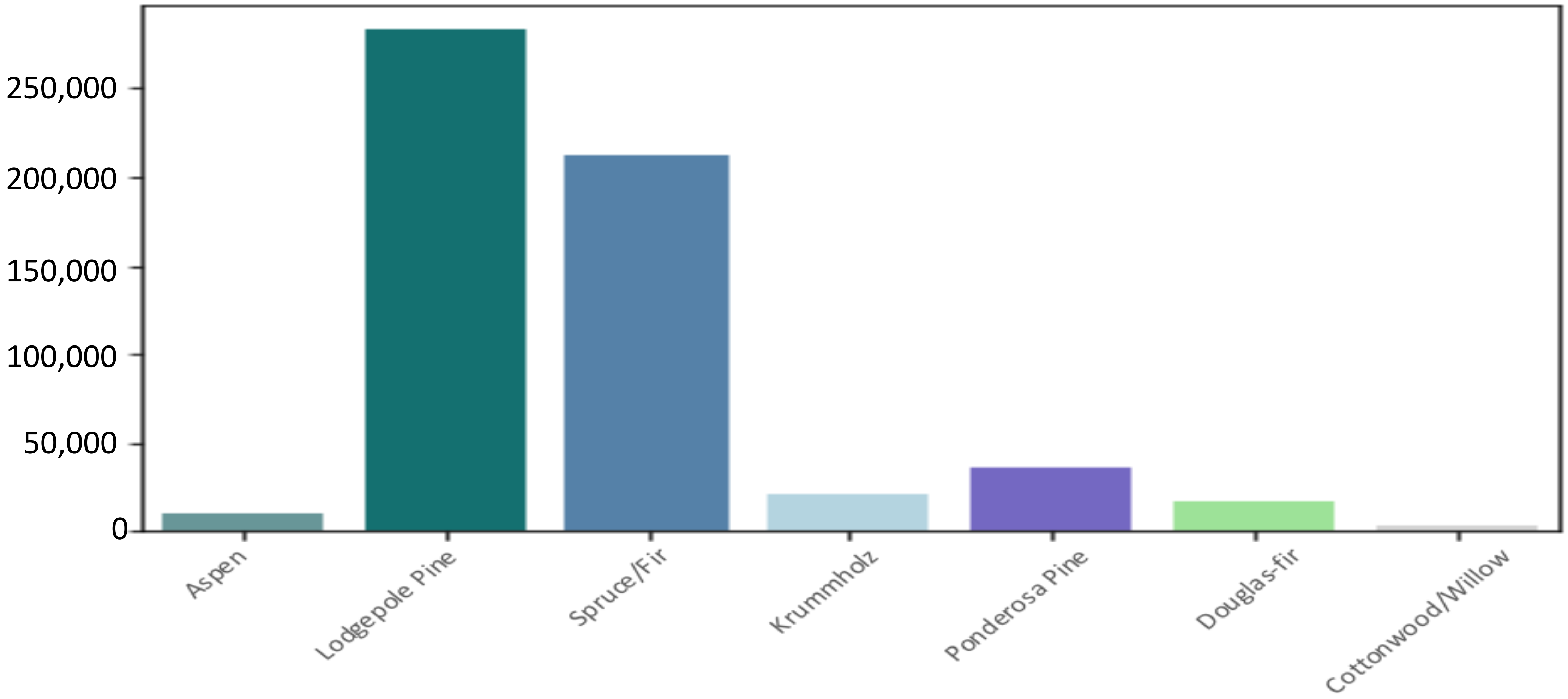


	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade_Noon	Hillshade_3pm	Horizontal_Distance_To_Fire_Points
Elevation	1	0.016	-0.24	0.31	0.093	0.37	0.11	0.21	0.059	0.15
Aspect	0.016	1	0.079	0.017	0.07	0.025	-0.58	0.34	0.65	-0.11
Slope	-0.24	0.079	1	-0.011	0.27	-0.22	-0.33	-0.53	-0.18	-0.19
Horizontal_Distance_To_Hydrology	0.31	0.017	-0.011	1	0.61	0.072	-0.027	0.047	0.052	0.052
Vertical_Distance_To_Hydrology	0.093	0.07	0.27	0.61	1	-0.046	-0.17	-0.11	0.035	-0.07
Horizontal_Distance_To_Roadways	0.37	0.025	-0.22	0.072	-0.046	1	0.034	0.19	0.11	0.33
Hillshade_9am	0.11	-0.58	-0.33	-0.027	-0.17	0.034	1	0.01	-0.78	0.13
Hillshade_Noon	0.21	0.34	-0.53	0.047	-0.11	0.19	0.01	1	0.59	0.057
Hillshade_3pm	0.059	0.65	-0.18	0.052	0.035	0.11	-0.78	0.59	1	-0.048
Horizontal_Distance_To_Fire_Points	0.15	-0.11	-0.19	0.052	-0.07	0.33	0.13	0.057	-0.048	1

Exploration

Forest Cover Types

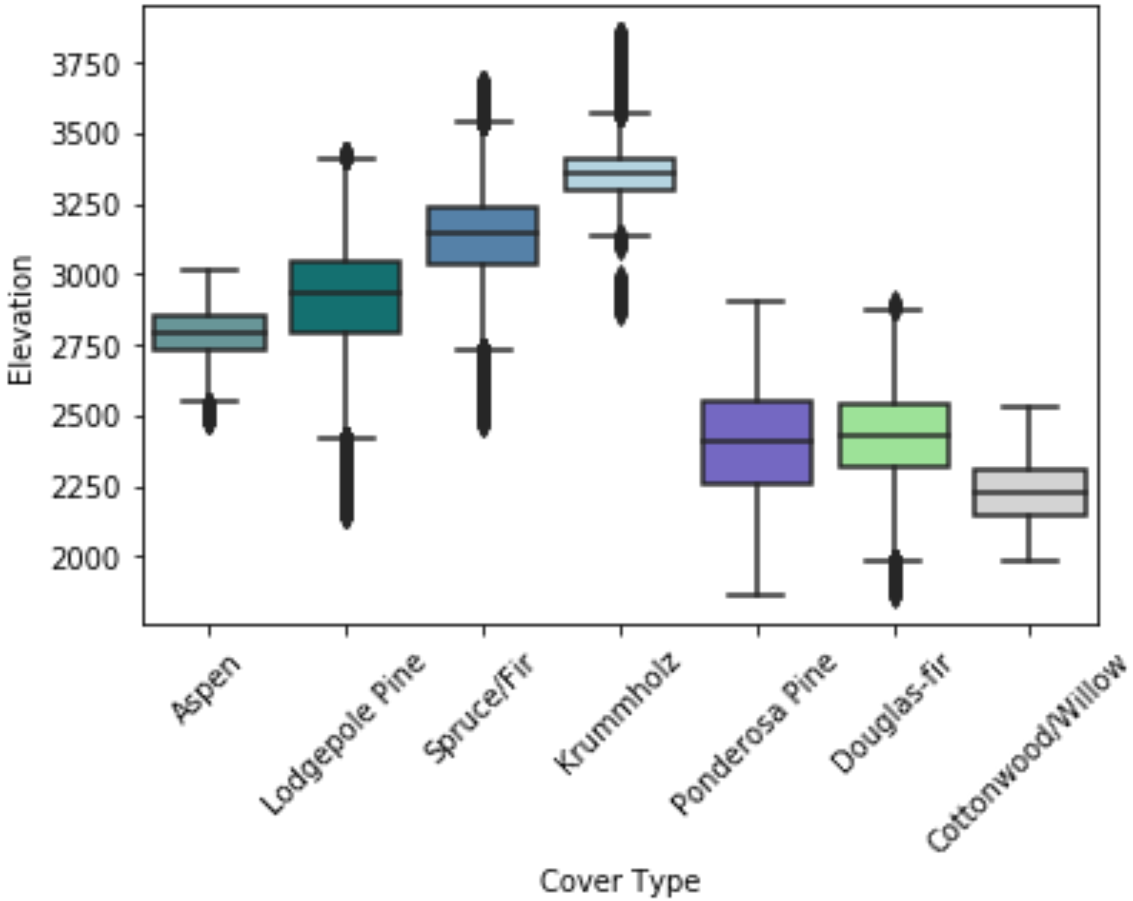
Frequency of Forest Cover Types



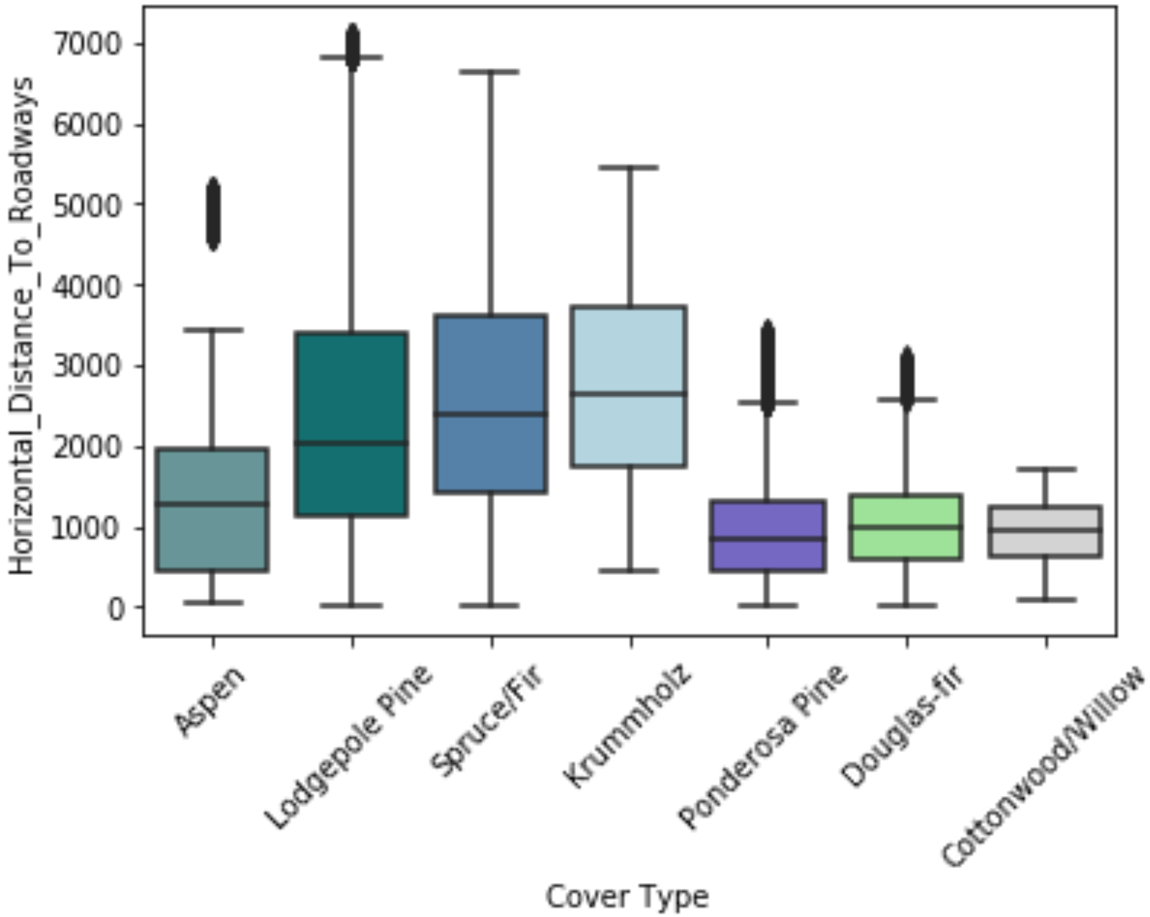
Exploration

Elevation & Horizontal Distance to Nearest Roadways

Elevation (meters)

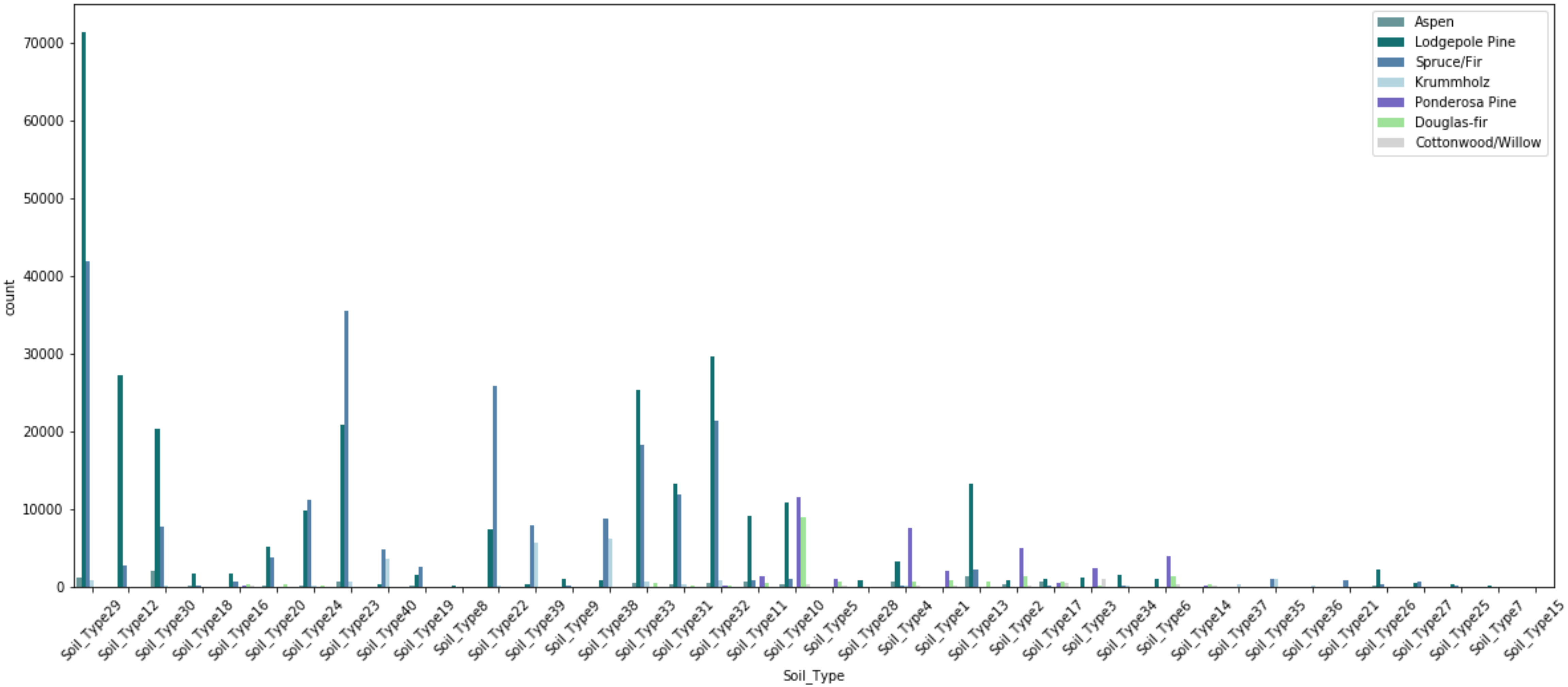


Horizontal Distance to Nearest Roadway



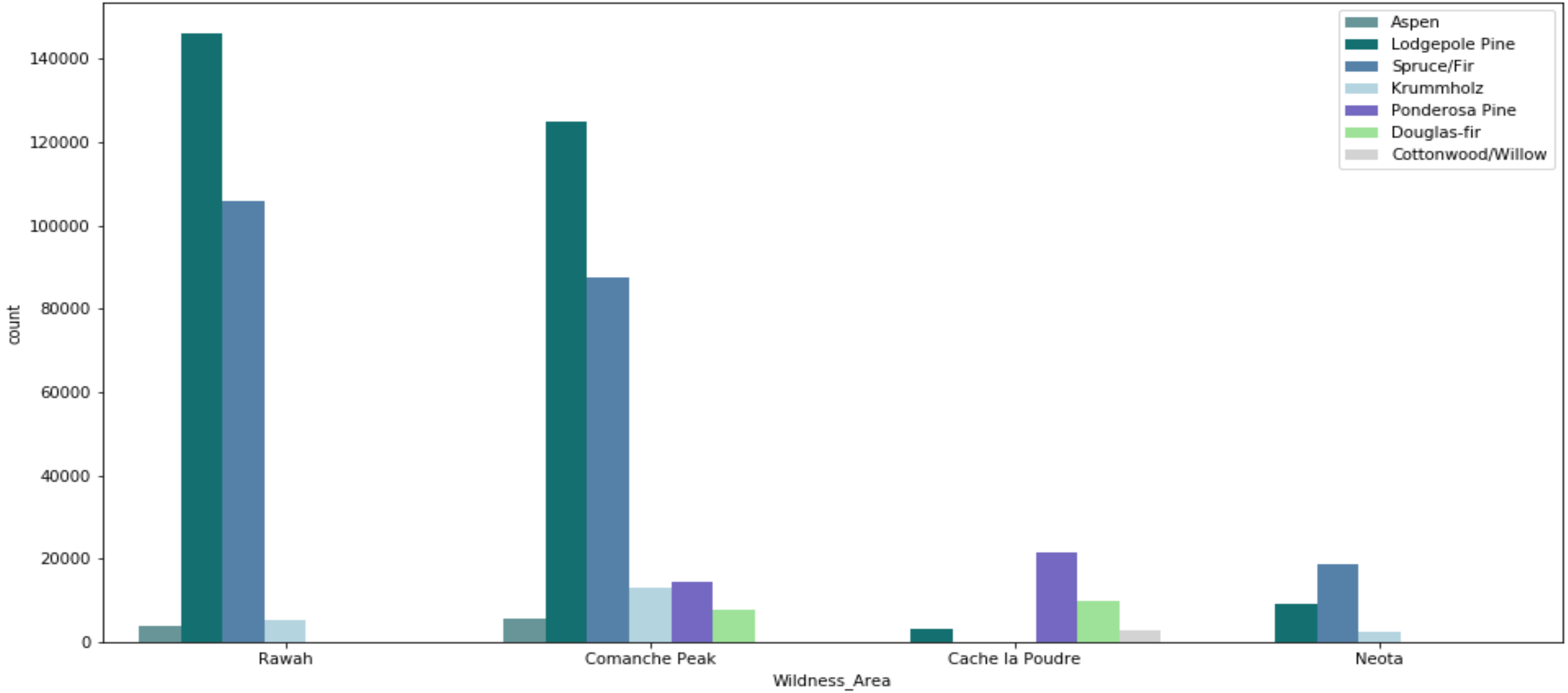
Exploration

Soil Types



Exploration

Wilderness Areas



Machine Learning Classifier

Performance: Metrics

Model	METRICS			
	Accuracy	Precision	Recall	F-1 Score
K-Nearest Neighbors	0.879	0.877	0.879	0.878
Random Forest	0.912	0.911	0.912	0.911
XGBoost	0.918	0.917	0.918	0.917

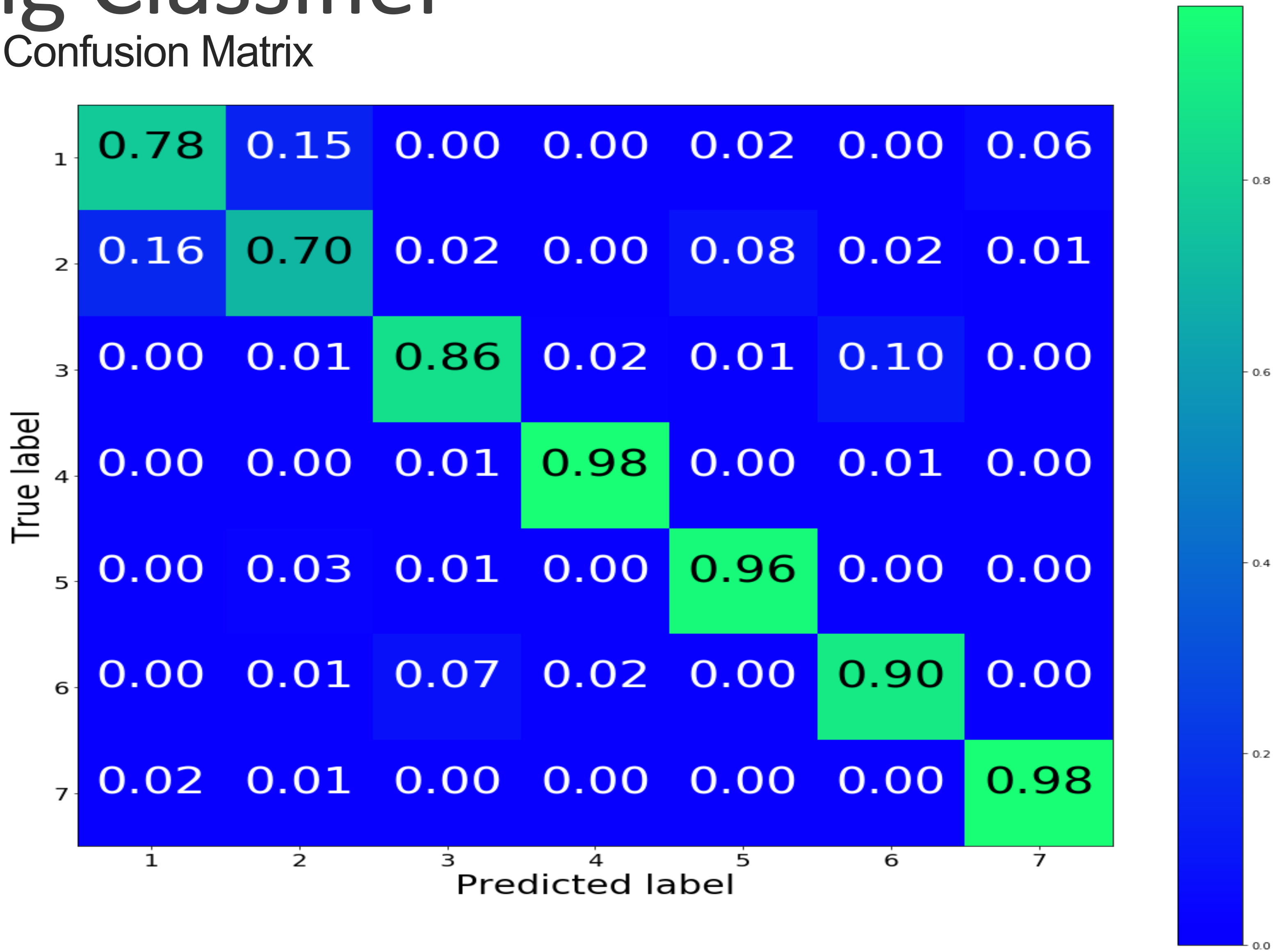
Machine Learning Classifier

Performance: K-Nearest Neighbors Confusion Matrix

Misclassification Findings:

0.15: Spruce/Fir with Lodgepole Pine
0.16: Lodgepole Pine with Spruce/Fir
0.08: Lodgepole Pine with Aspen
0.10: Ponderosa Pine with Douglas-fir
0.07: Douglas-fir with Ponderosa Pine

MATRIX KEY	
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz



	1	2	3	4	5	6	7
1	0.78	0.15	0.00	0.00	0.02	0.00	0.06
2	0.16	0.70	0.02	0.00	0.08	0.02	0.01
3	0.00	0.01	0.86	0.02	0.01	0.10	0.00
4	0.00	0.00	0.01	0.98	0.00	0.01	0.00
5	0.00	0.03	0.01	0.00	0.96	0.00	0.00
6	0.00	0.01	0.07	0.02	0.00	0.90	0.00
7	0.02	0.01	0.00	0.00	0.00	0.00	0.98

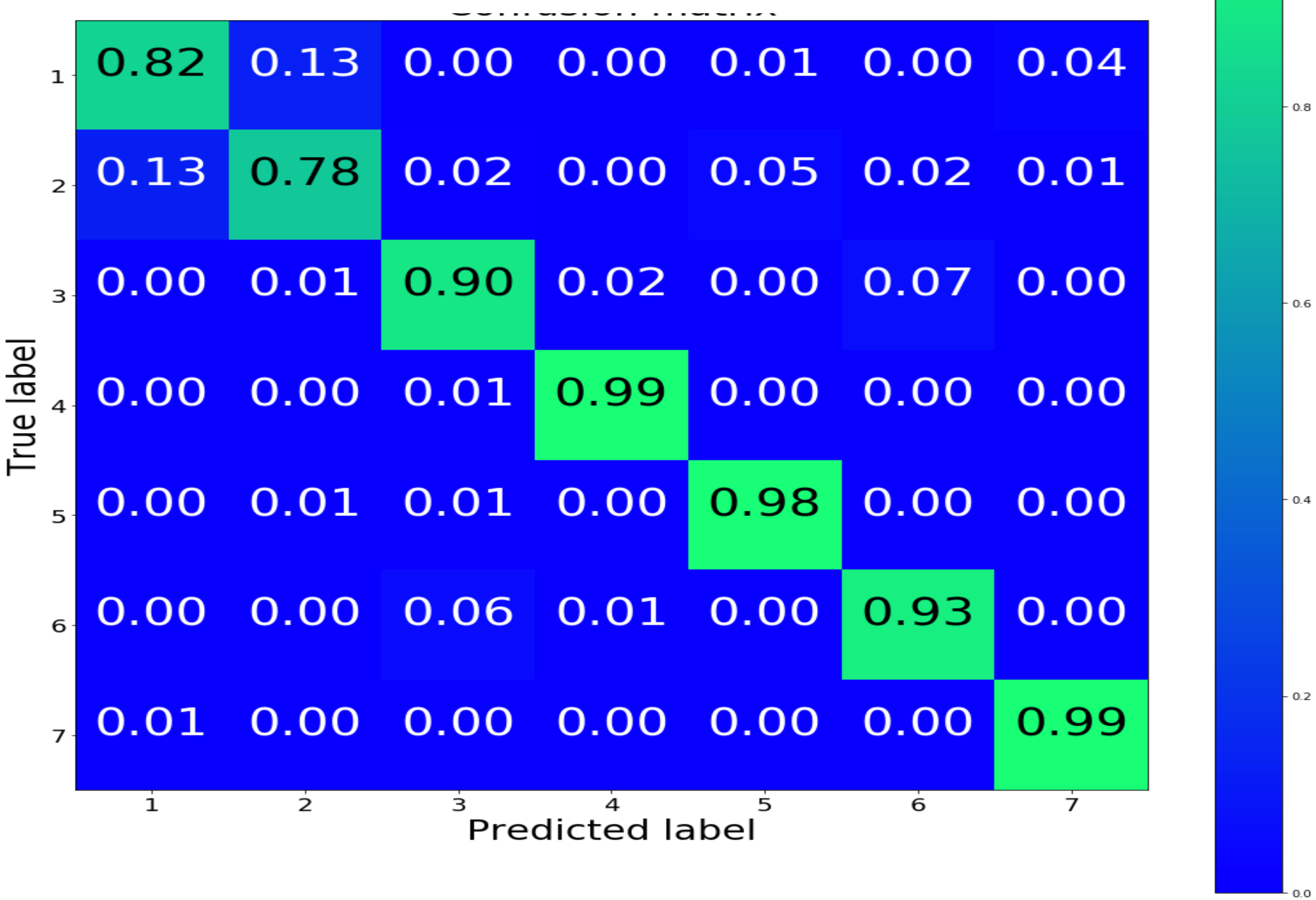
Machine Learning Classifier

Performance: Random Forest Confusion Matrix

Misclassification Findings:

0.13: Spruce/Fir with Lodgepole Pine
0.13: Lodgepole Pine with Spruce/Fir
0.05: Lodgepole Pine with Aspen
0.07: Ponderosa Pine with Douglas-fir
0.06: Douglas-fir with Ponderosa Pine

MATRIX KEY	
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz



	1	2	3	4	5	6	7
1	0.82	0.13	0.00	0.00	0.01	0.00	0.04
2	0.13	0.78	0.02	0.00	0.05	0.02	0.01
3	0.00	0.01	0.90	0.02	0.00	0.07	0.00
4	0.00	0.00	0.01	0.99	0.00	0.00	0.00
5	0.00	0.01	0.01	0.00	0.98	0.00	0.00
6	0.00	0.00	0.06	0.01	0.00	0.93	0.00
7	0.01	0.00	0.00	0.00	0.00	0.00	0.99

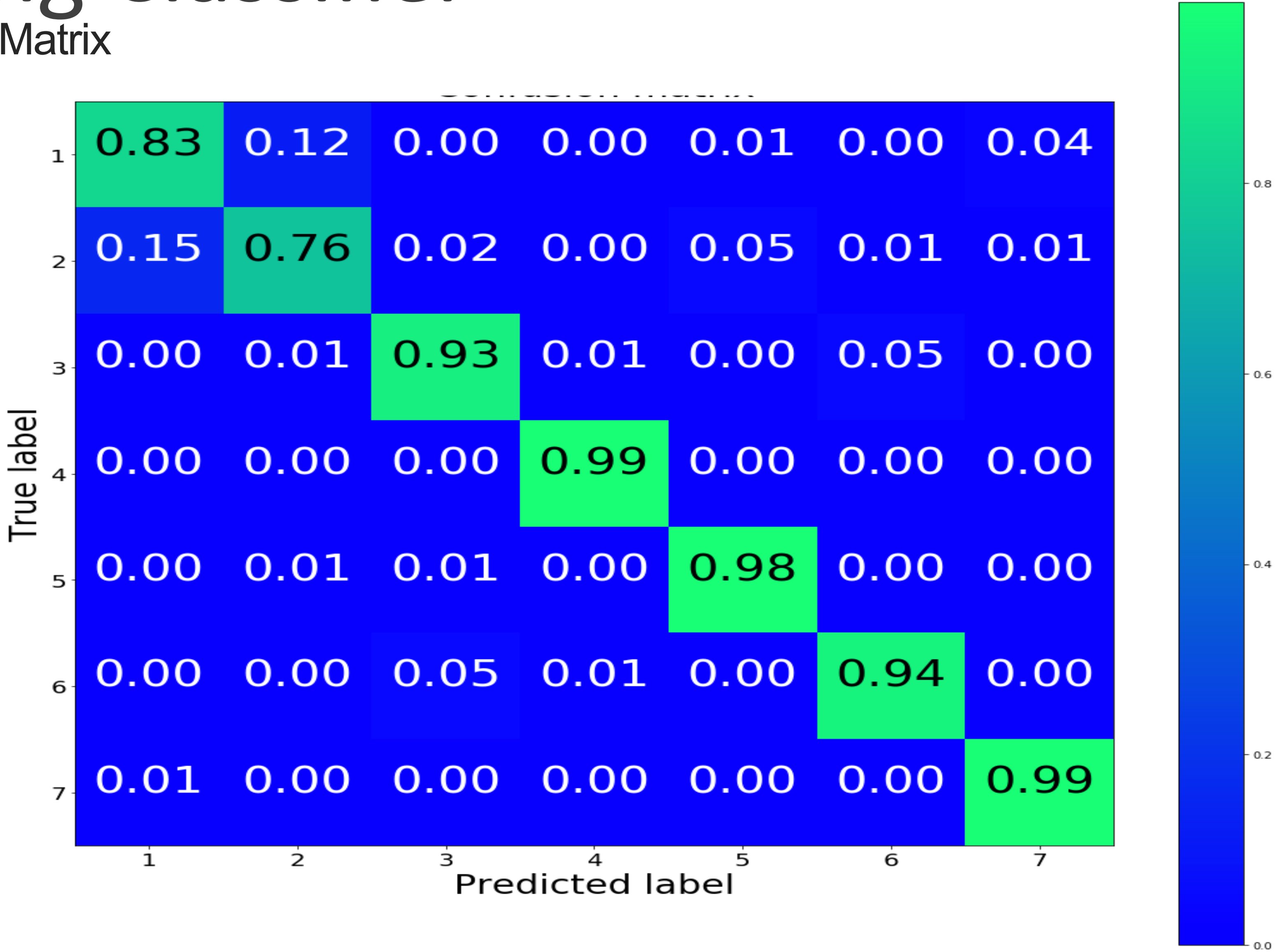
Machine Learning Classifier

Performance: XGBoost Confusion Matrix

Misclassification Findings:

0.15: Spruce/Fir with Lodgepole Pine
0.12: Lodgepole Pine with Spruce/Fir
0.05: Lodgepole Pine with Aspen
0.05: Ponderosa Pine with Douglas-fir
0.05: Douglas-fir with Ponderosa Pine

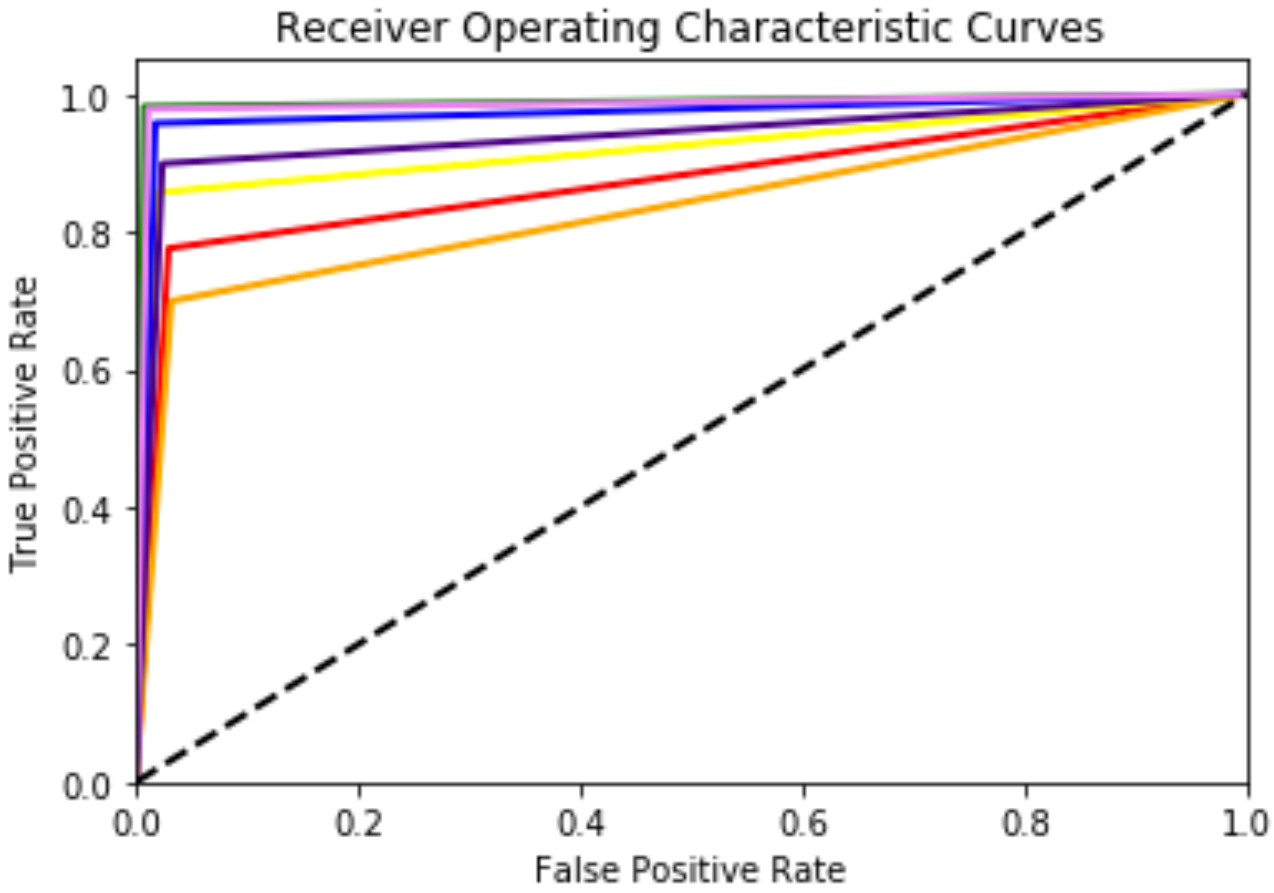
MATRIX KEY	
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz



True label \ Predicted label	1	2	3	4	5	6	7
1	0.83	0.12	0.00	0.00	0.01	0.00	0.04
2	0.15	0.76	0.02	0.00	0.05	0.01	0.01
3	0.00	0.01	0.93	0.01	0.00	0.05	0.00
4	0.00	0.00	0.00	0.99	0.00	0.00	0.00
5	0.00	0.01	0.01	0.00	0.98	0.00	0.00
6	0.00	0.00	0.05	0.01	0.00	0.94	0.00
7	0.01	0.00	0.00	0.00	0.00	0.00	0.99

Machine Learning Classifier

Performance: K-Nearest Neighbors ROC Curve

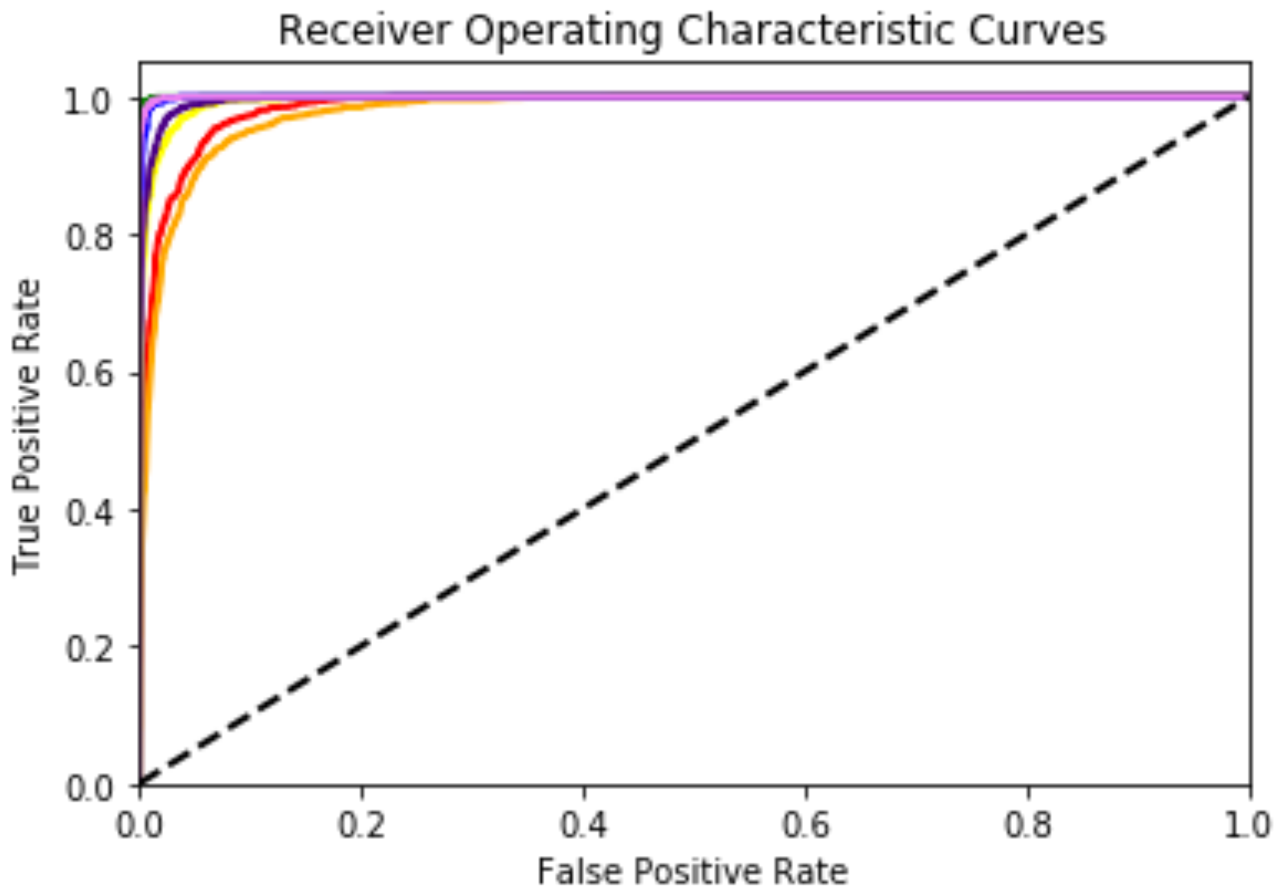


Area Under the Curve (AUC)

🟥	Spruce/Fir:	0.873
🟧	Lodgepole Pine:	0.834
🟨	Ponderosa Pine:	0.919
🟩	Cottonwood/Willow:	0.987
🟦	Aspen:	0.970
🟪	Douglas-fir:	0.938
🟪	Krummholz:	0.984

Machine Learning Classifier

Performance: Random Forest ROC Curve

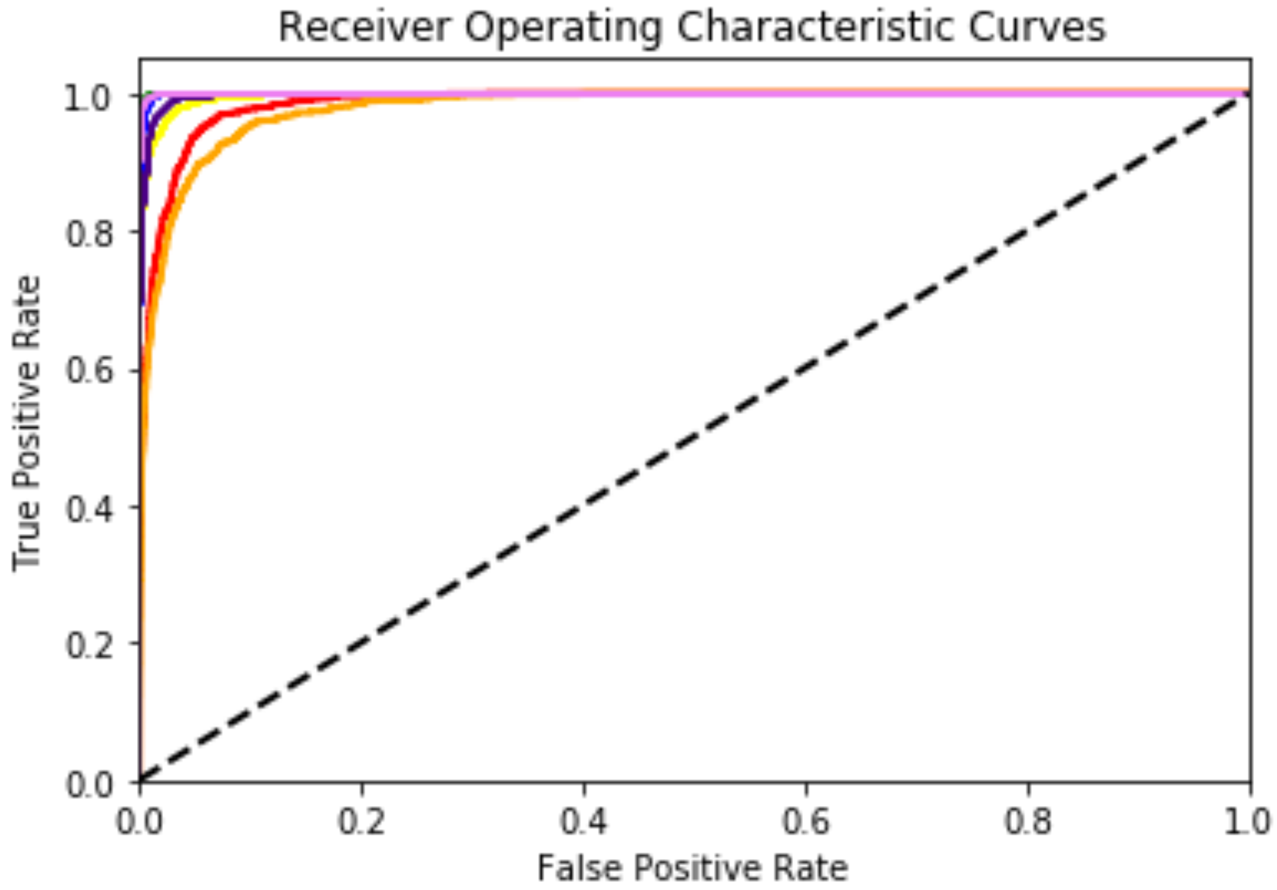


Area Under the Curve (AUC)

🟥	Spruce/Fir:	0.985
🟧	Lodgepole Pine:	0.978
🟨	Ponderosa Pine:	0.994
🟩	Cottonwood/Willow:	0.999
🟦	Aspen:	0.999
🟪	Douglas-fir:	0.996
🟪	Krummholz:	0.999

Machine Learning Classifier

Performance: XGBoost ROC Curve



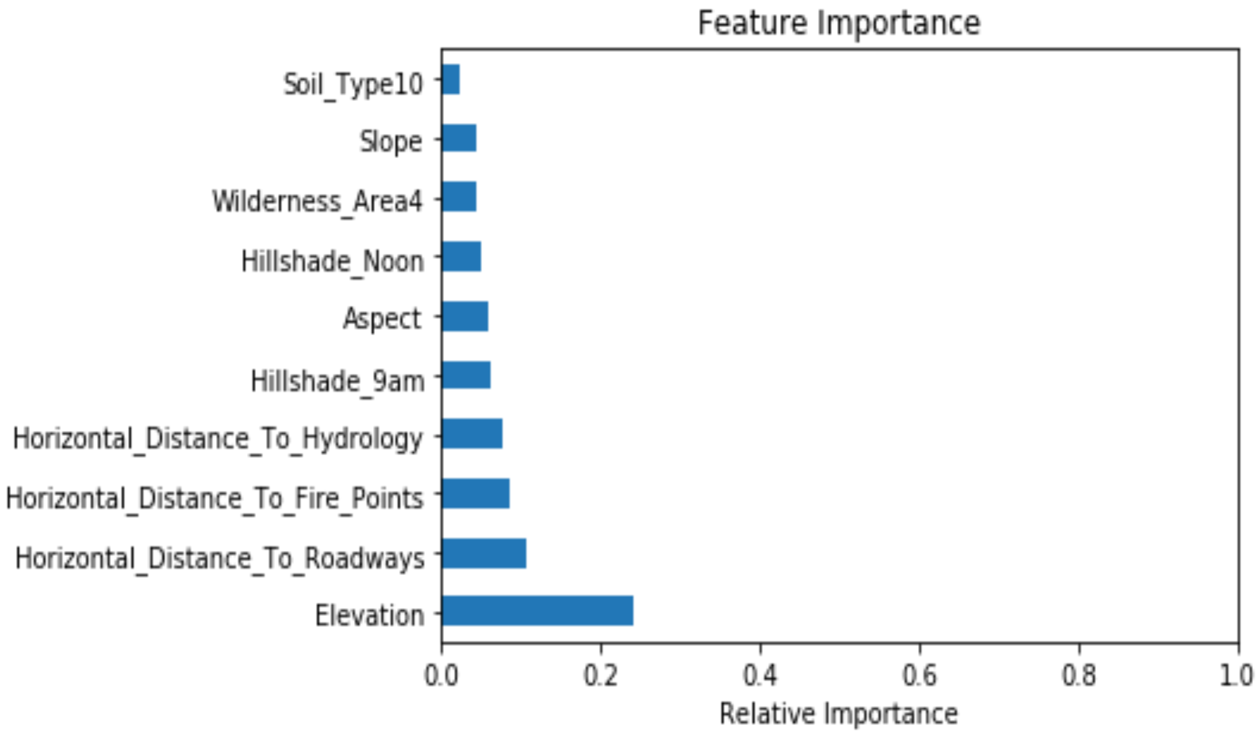
Area Under the Curve (AUC)

🟥	Spruce/Fir:	0.986
🟧	Lodgepole Pine:	0.979
🟨	Ponderosa Pine:	0.996
🟩	Cottonwood/Willow:	0.999
🟦	Aspen:	0.999
🟪	Douglas-fir:	0.997
🟪	Krummholz:	0.999

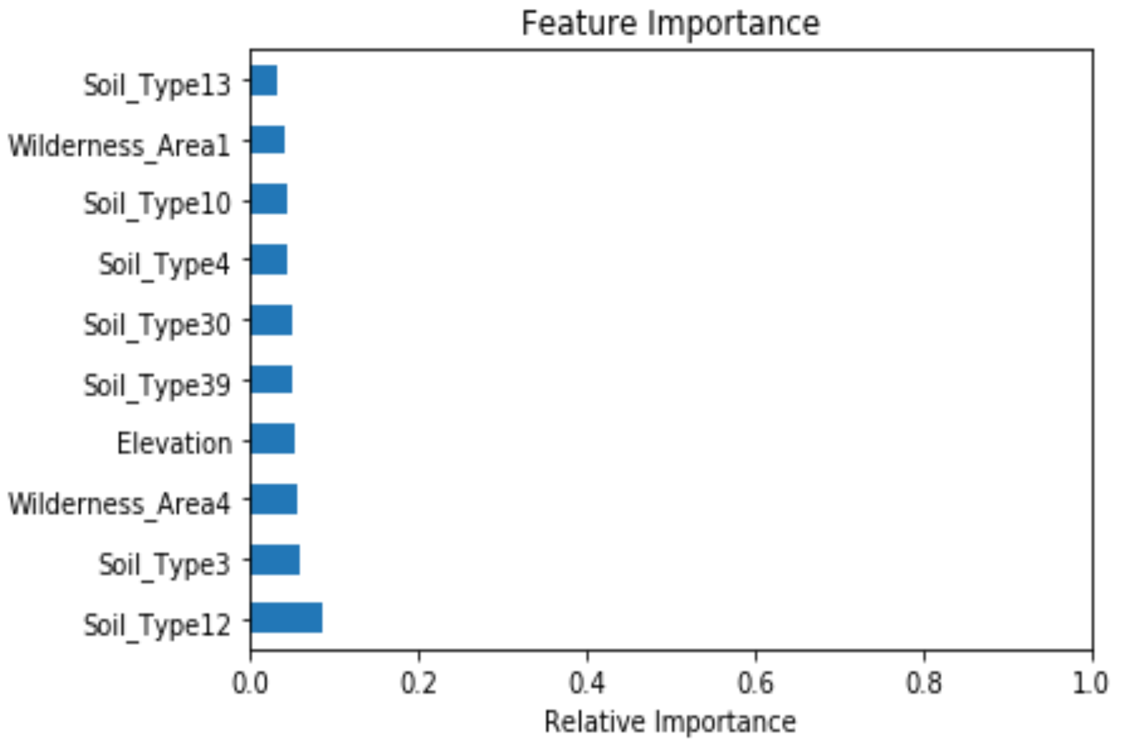
Machine Learning Classifier

Performance: Feature Importance

Random Forest



XGBoost



Future Work & Next Steps

Future Work & Next Steps

- Perform additional analysis to determine predictions for wildfires, and assisting in their prevention
    - Obtain data on the history of wildfires within the Roosevelt National Forest

- Further tune the XGBoost model to improve accuracy and even out feature importance

- Determine if the XGBoost model will perform well for other forests within Colorado
    - Verify if additional forest cover type data has been collected by the College of Natural Resources at Colorado State University

Thank You!

Questions?

Appendix

Appendix: Data

Forest Cover Type Designation:

1 - Spruce/Fir
2 - Lodgepole Pine
3 - Ponderosa Pine
4 - Cottonwood/Willow
5 - Aspen
6 - Douglas-fir
7 - Krummholz

Wilderness Area Designation:

1 - Rawah Wilderness Area
2 - Neota Wilderness Area
3 - Comanche Peak Wilderness Area
4 - Cache la Poudre Wilderness Area

Appendix: Data

Soil Type Designation:

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.
5 Vanet family - Rock outcrop complex complex, rubbly.
6 Vanet - Wetmore families - Rock outcrop complex, stony.
7 Gothic family.
8 Supervisor - Limber families complex.
9 Troutville family, very stony.
10 Bullwark - Catamount families - Rock outcrop complex, rubbly.
11 Bullwark - Catamount families - Rock land complex, rubbly.
12 Legault family - Rock land complex, stony.
13 Catamount family - Rock land - Bullwark family complex, rubbly.
14 Pachic Argiborolis - Aquolis complex.
15 unspecified in the USFS Soil and ELU Survey.
16 Cryaquolis - Cryoborolis complex.
17 Gateview family - Cryaquolis complex.
18 Rogert family, very stony.
19 Typic Cryaquolis - Borohemists complex.
20 Typic Cryaquepts - Typic Cryaquolls complex.
21 Typic Cryaquolls - Leighcan family, till substratum complex.
22 Leighcan family, till substratum, extremely bouldery.
23 Leighcan family, till substratum - Typic Cryaquolls complex.
24 Leighcan family, extremely stony.
25 Leighcan family, warm, extremely stony.
26 Granile - Catamount families complex, very stony.
27 Leighcan family, warm - Rock outcrop complex, extremely stony.
28 Leighcan family - Rock outcrop complex, extremely stony.
29 Como - Legault families complex, extremely stony.
30 Como family - Rock land - Legault family complex, extremely stony.
31 Leighcan - Catamount families complex, extremely stony.
32 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
33 Leighcan - Catamount families - Rock outcrop complex, extremely stony.
34 Cryorthents - Rock land complex, extremely stony.
35 Cryumbrepts - Rock outcrop - Cryaquepts complex.
36 Bross family - Rock land - Cryumbrepts complex, extremely stony.
37 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
38 Leighcan - Moran families - Cryaquolls complex, extremely stony.
39 Moran family - Cryorthents - Leighcan family complex, extremely stony.
40 Moran family - Cryorthents - Rock land complex, extremely stony.

Appendix: Notebook Link

- GitHub Link: Predicting Forest Cover Types