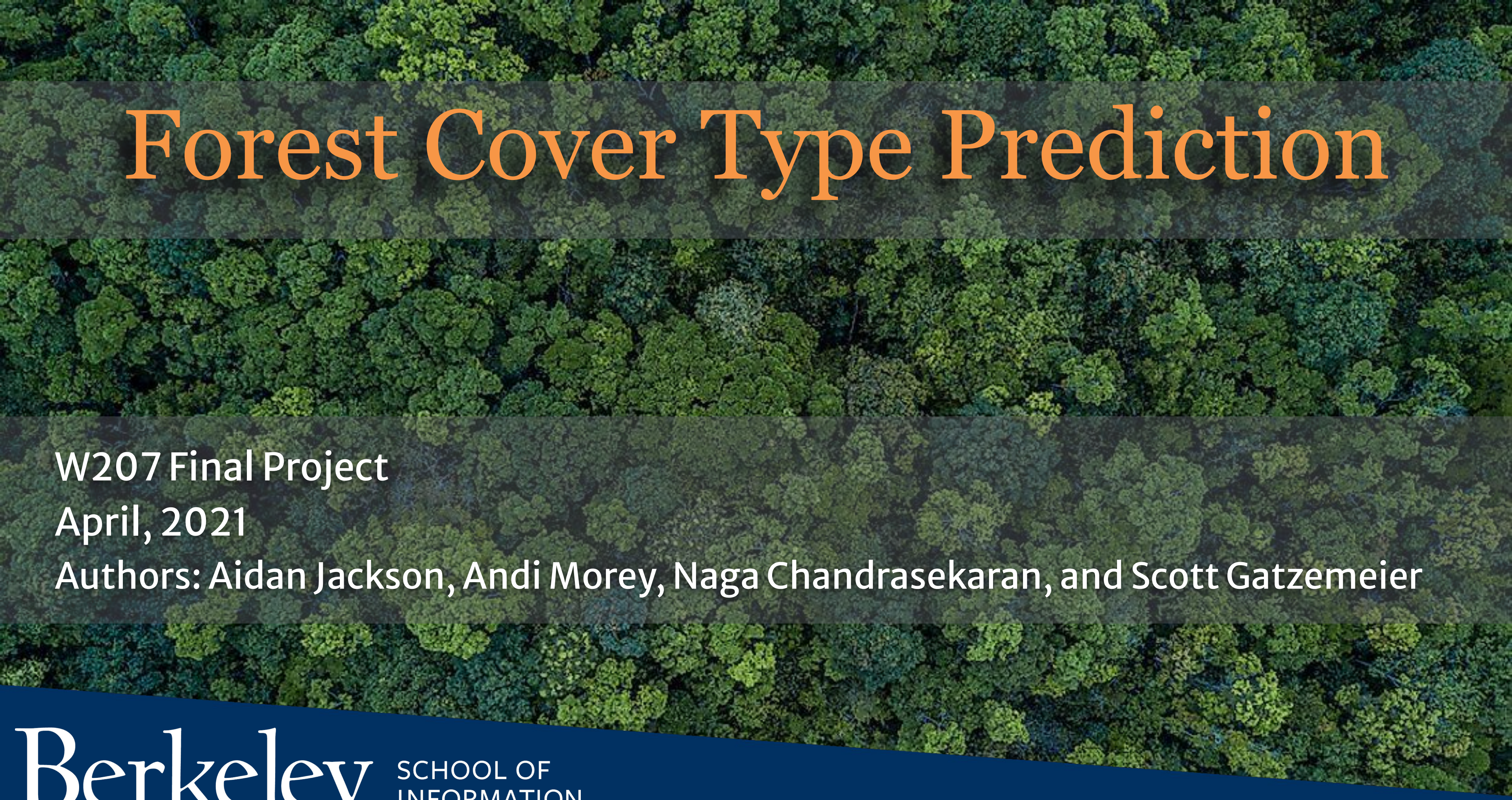


Forest Cover Type Prediction

W207 Final Project

April, 2021

Authors: Aidan Jackson, Andi Morey, Naga Chandrasekaran, and Scott Gatzemeier

Berkeley SCHOOL OF INFORMATION

Agenda

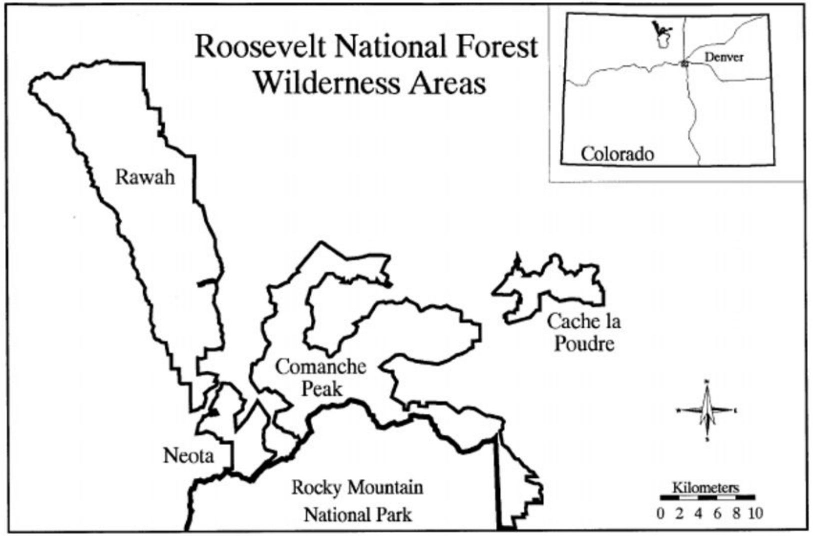
1. Exploratory Data Analysis
2. Clean/Format Data and Feature Engineering
3. Initial Machine Learning Models
4. Hyperparameter Tuning
5. Evaluation of the Best Model
6. Interpret Model Results
7. Summary & Conclusions

Problem Statement

Cover Types

Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

Areas



Classify the cover type for areas

Dataset Information

Data

- 581,012 instances
- 565,892 test set
- 15, 120 training set

Attributes

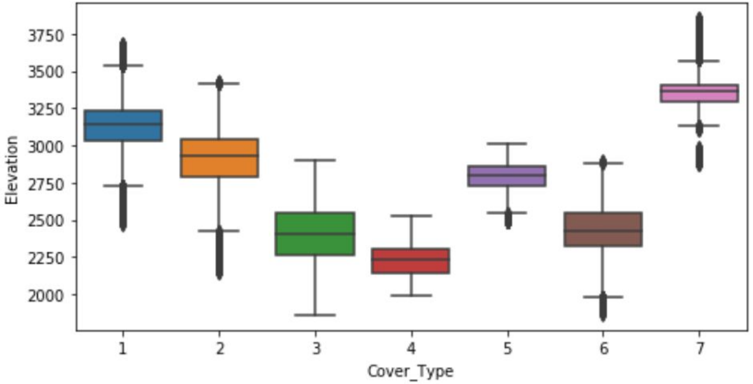
- Numeric, Categorical
- 12 categories
- 54 total columns
- 44 data columns binary

Attribute Details

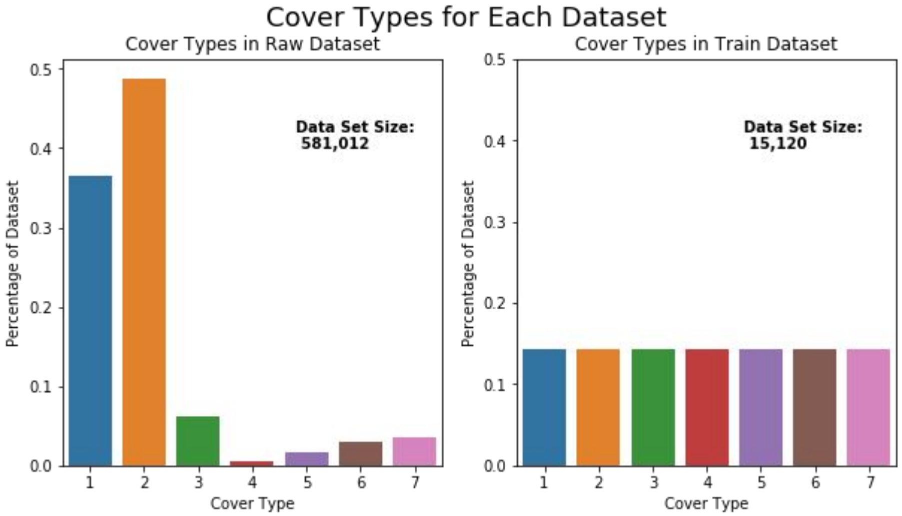
1. Elevation
2. Aspect
3. Slope
4. Horizontal distance to hydrology
5. Vertical distance to hydrology
6. Horizontal distance to roadways
7. Horizontal distance to firepoints
8. Hill shade 9am: RBG representation
9. Hill shade noon: RBG representation
10. Hill shade 3pm: RBG representation
11. Wilderness areas: 4 binary data
12. Soil type: 40 binary data

Exploratory Data Analysis

Cover Type Key Important Feature

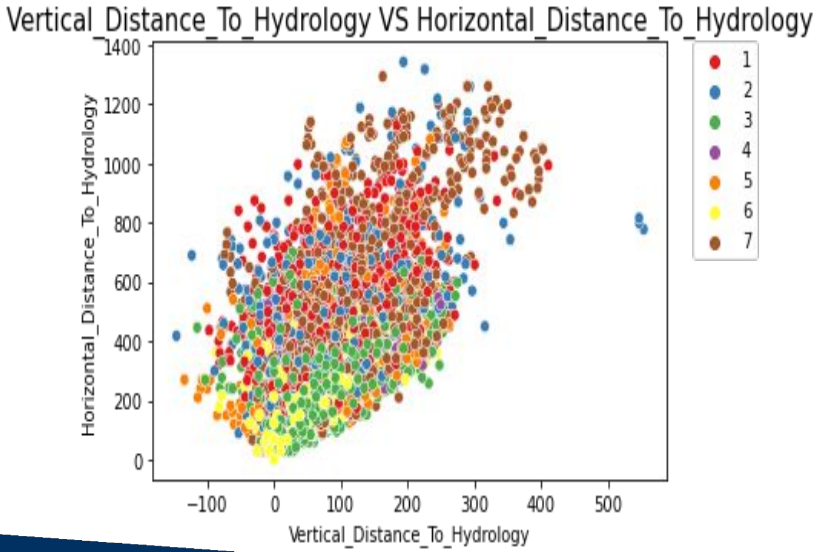


Cover Type Distributions

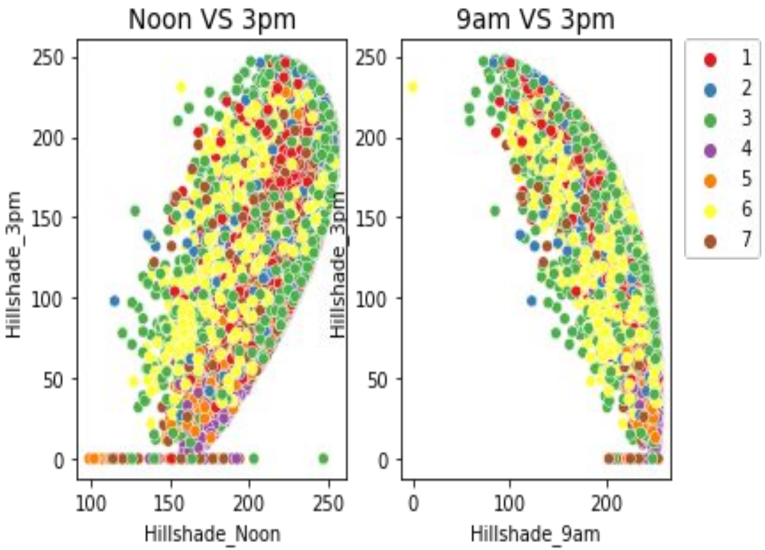


Exploratory Data Analysis

Training Data Frame



Hillshade Correlation

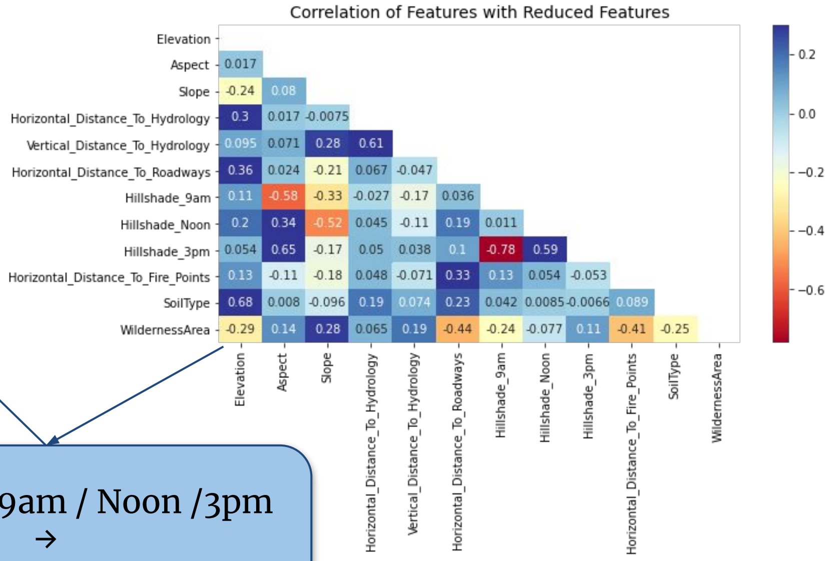


Berkeley SCHOOL OF INFORMATION

Feature Engineering

Original Dataset: 54 Parameters

Soil Type and Wilderness Area Reduce 44 binary to 2 categorical parameters


Correlation of Features with Reduced 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	SoilType	WildernessArea
Elevation												
Aspect	0.017											
Slope	-0.24	0.08										
Horizontal_Distance_To_Hydrology	0.3	0.017	-0.0075									
Vertical_Distance_To_Hydrology	0.095	0.071	0.28	0.61								
Horizontal_Distance_To_Roadways	0.36	0.024	-0.21	0.067	-0.047							
Hillshade_9am	0.11	-0.58	-0.33	-0.027	-0.17	0.036						
Hillshade_Noon	0.2	0.34	-0.52	0.045	-0.11	0.19	0.011					
Hillshade_3pm	0.054	0.65	-0.17	0.05	0.038	0.1	-0.78	0.59				
Horizontal_Distance_To_Fire_Points	0.13	-0.11	-0.18	0.048	-0.071	0.33	0.13	0.054	-0.053			
SoilType	0.68	0.008	-0.096	0.19	0.074	0.23	0.042	0.0085	-0.0066	0.089		
WildernessArea	-0.29	0.14	0.28	0.065	0.19	-0.44	-0.24	-0.077	0.11	-0.41	-0.25	

Hillshade_9am / Noon /3pm → Average_Hillshade

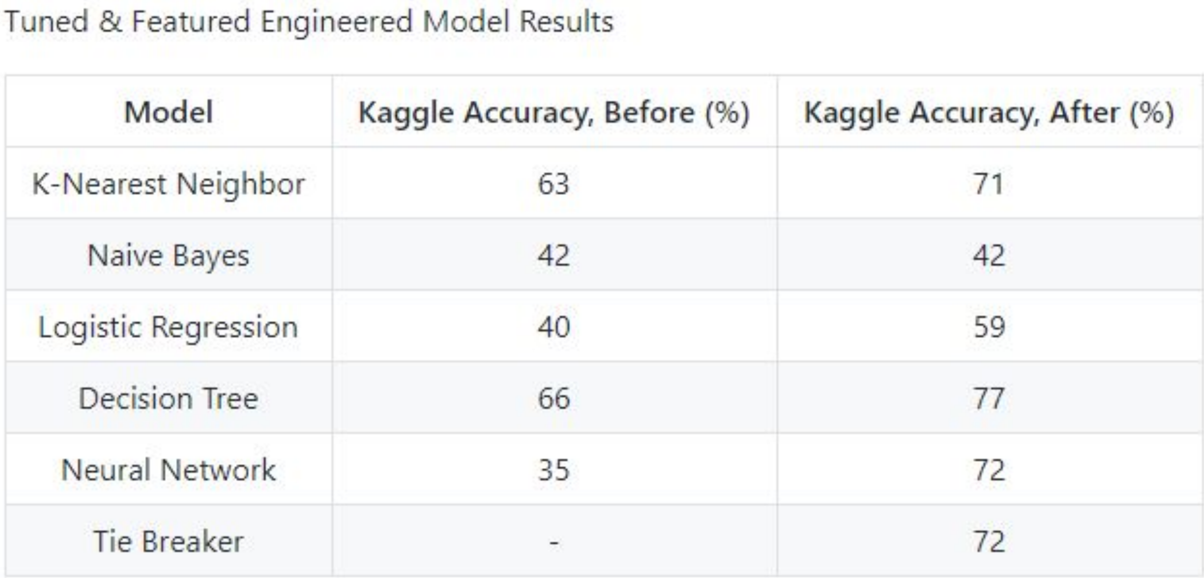
Horizontal / Vertical Distance to Hydrology → Total Euclidean Distance

Normalization of all Distance values

Transformation of Aspect to Care only of N/S Facing Topos

Berkeley SCHOOL OF INFORMATION

Machine Learning Model Progression



Tuned & Featured Engineered Model Results

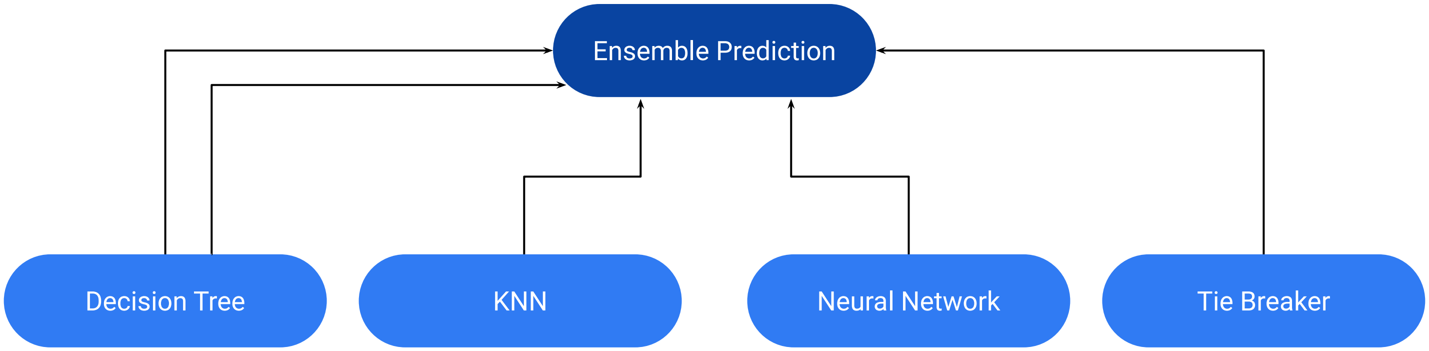
Model	Kaggle Accuracy, Before (%)	Kaggle Accuracy, After (%)
K-Nearest Neighbor	63	71
Naive Bayes	42	42
Logistic Regression	40	59
Decision Tree	66	77
Neural Network	35	72
Tie Breaker	-	72

Hyperparameter Tuning

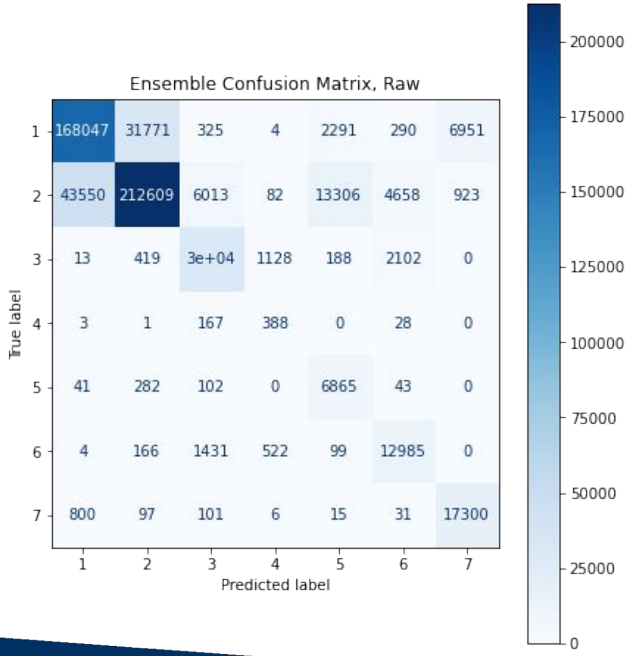
- Naive Bayes and Logistic Regression discarded due to low accuracy
- Random Forest had best individual performance
- KNN: K = 1, Euclidean distance
- NN: 9 hidden layers, 100 nodes each

Evaluation of the Best Model

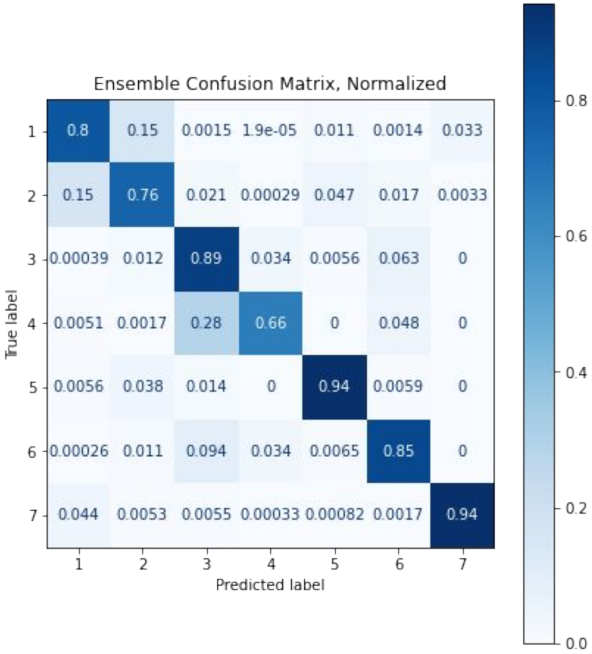
- Best accuracy: 79.579%
- Position 197 / 1693 on the Leaderboard



Interpret Model Results



	1	2	3	4	5	6	7
1	168047	31771	325	4	2291	290	6951
2	43550	212609	6013	82	13306	4658	923
3	13	419	3e+04	1128	188	2102	0
4	3	1	167	388	0	28	0
5	41	282	102	0	6865	43	0
6	4	166	1431	522	99	12985	0
7	800	97	101	6	15	31	17300

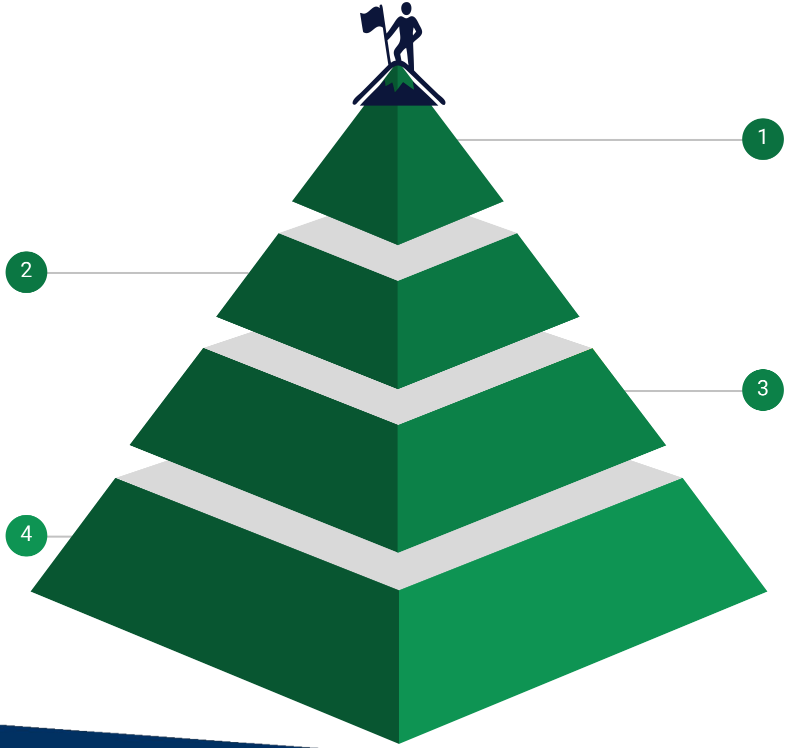


	1	2	3	4	5	6	7
1	0.8	0.15	0.0015	1.9e-05	0.011	0.0014	0.033
2	0.15	0.76	0.021	0.00029	0.047	0.017	0.0033
3	0.00039	0.012	0.89	0.034	0.0056	0.063	0
4	0.0051	0.0017	0.28	0.66	0	0.048	0
5	0.0056	0.038	0.014	0	0.94	0.0059	0
6	0.00026	0.011	0.094	0.034	0.0065	0.85	0
7	0.044	0.0053	0.0055	0.00033	0.00082	0.0017	0.94

Summary & Conclusions

- **Ensemble Model Achieved Best Results**
  - Power of Group Consensus
- **Cover Type 1 and 2 Dominate in Test Data**
  - Make up 85% of Data
- **Future Use Case Challenges**
  - Specific to Wilderness areas in Colorado
  - Data Collection Method Unknown

Lessons Learned



Last 10% was >90% of work

Perseverance is key, shocking how much didn’t work

Hyperparameter Tuning is Compute Intensive

Way too much time spent waiting for models to return results

Feature Engineering is difficult

- Deep understanding of data
- Model Specific
- Imbalanced Dataset is a challenge

Collaboration is Key

Good Communication and task alignment is critical to timely completion

Questions?