# Forest Cover Type Prediction

W207 Final Project – Mid Term Presentation

March, 2021

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

Berkeley SCHOOL OF INFORMATION

# Agenda to be deleted

- Speaker 1:
  - Into and first EDA Slide – Scott
- Speaker 2:
  - Additional EDA – Naga
- Speaker 3:
  - Preliminary Model Results – Andi
- Speaker 4:
  - Stuck Points and Next Steps – Aidan

# Overview

Problem Statement:

- Predict the predominant kind of tree cover from strictly cartographic variables
- Seven Classification types:
  - Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz

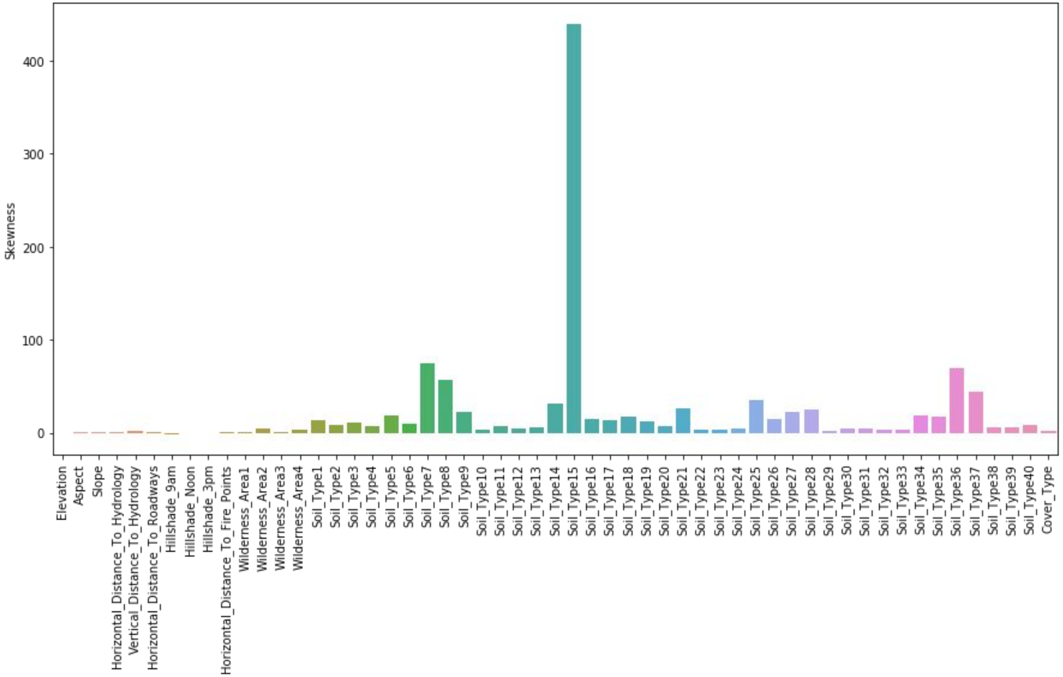
Data Set:

- Actual forest cover type determined by US Forest Service (USFS) for a 30 x 30 meter cell from Northern Colorado

# EDA - Dataset

## Dataset Analysis:

- Test dataset: 565892 observations with 55 features
- Training dataset: 15120 observations with 56 features, including cover type

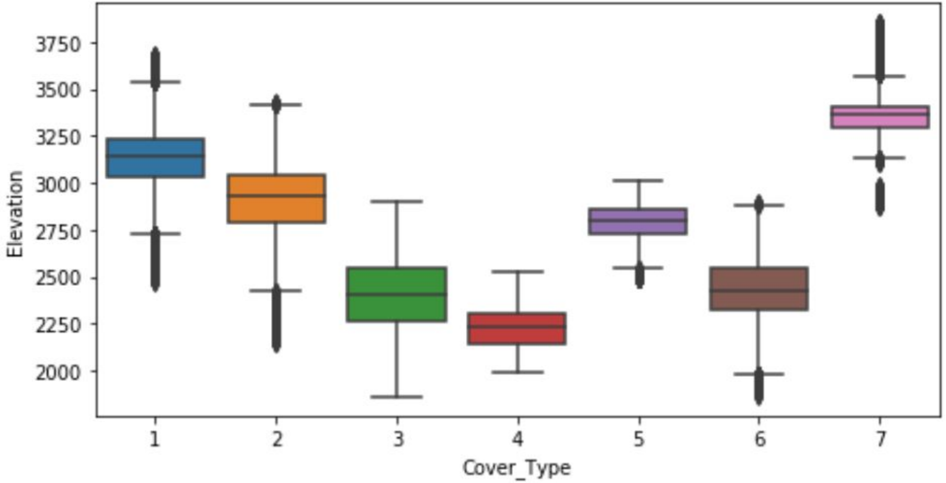


Berkeley SCHOOL OF INFORMATION

# EDA - Cover Type

Exploratory Data Analysis

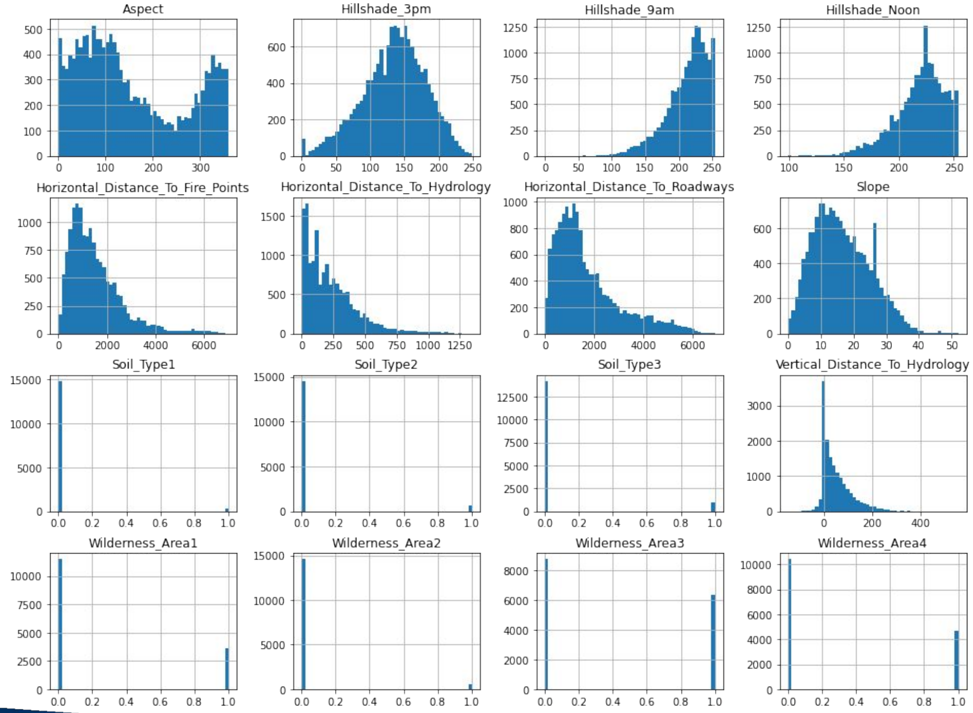
- Elevation has largest impact on cover type



# EDA - Features

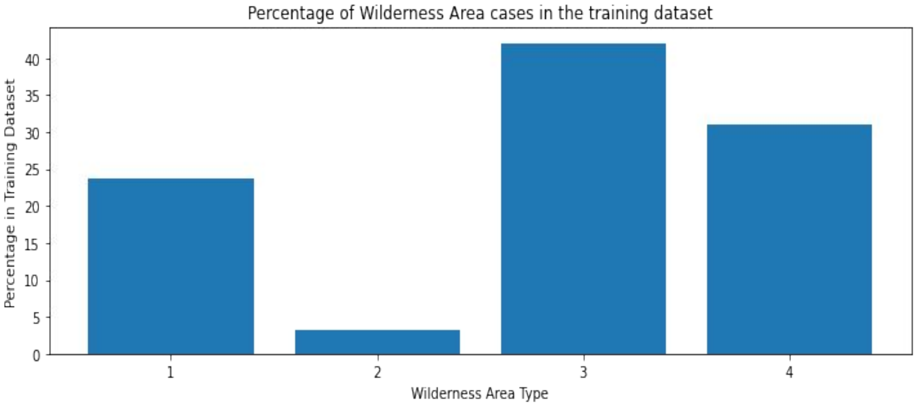
## Feature Dataset:

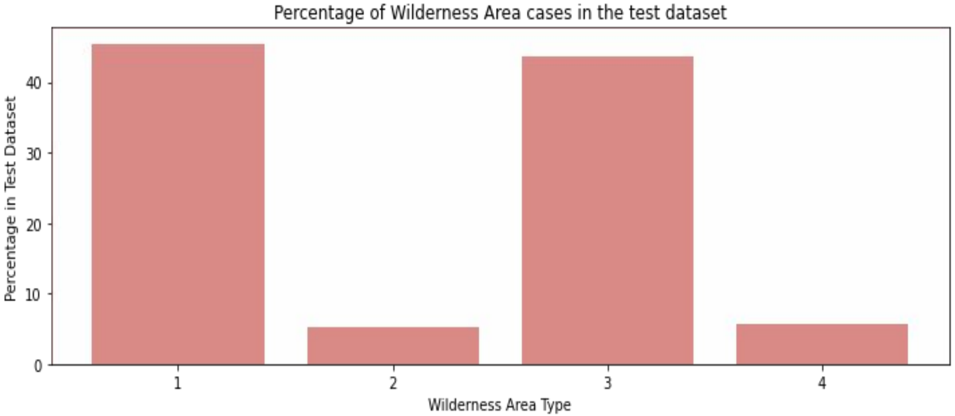
- Elevation
- Aspect – influence on temperature
- Hill Slope
- Distance to Water
- Distance to Roads
- Shade
- Distance to wildfire ignition points
- Wilderness Area
- Soil Type (40 binary columns)
- Cover type (7 designations)



Berkeley SCHOOL OF INFORMATION

# EDA - Wilderness Area Type

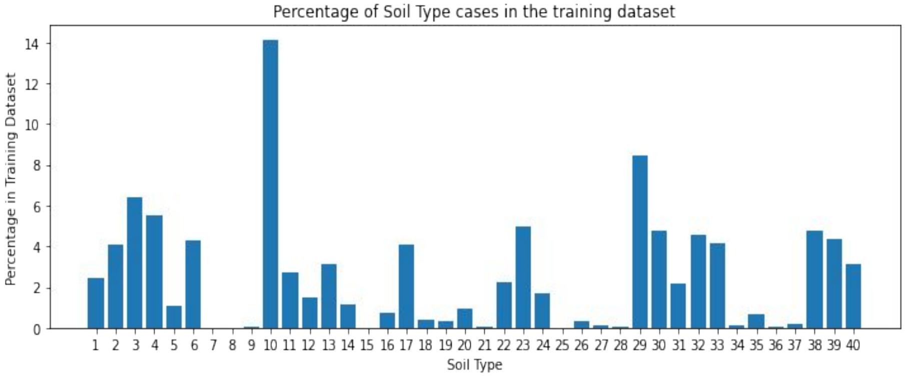


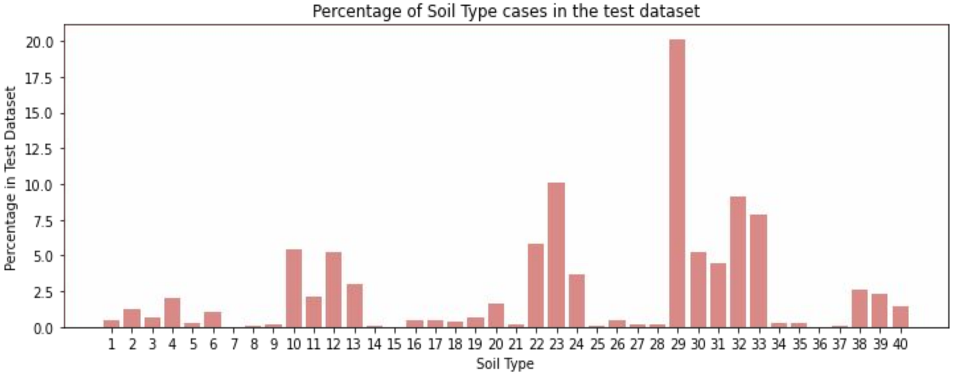


There are 4 wilderness area types: Rawah, Neota, Comanche, and Cache la Poudre

- Training data set: Areas 1, 3, and 4 are well represented
- Test data set: Areas 1 and 3 with high representation (areas 2 and 4 being very low)

# EDA - Soil Type





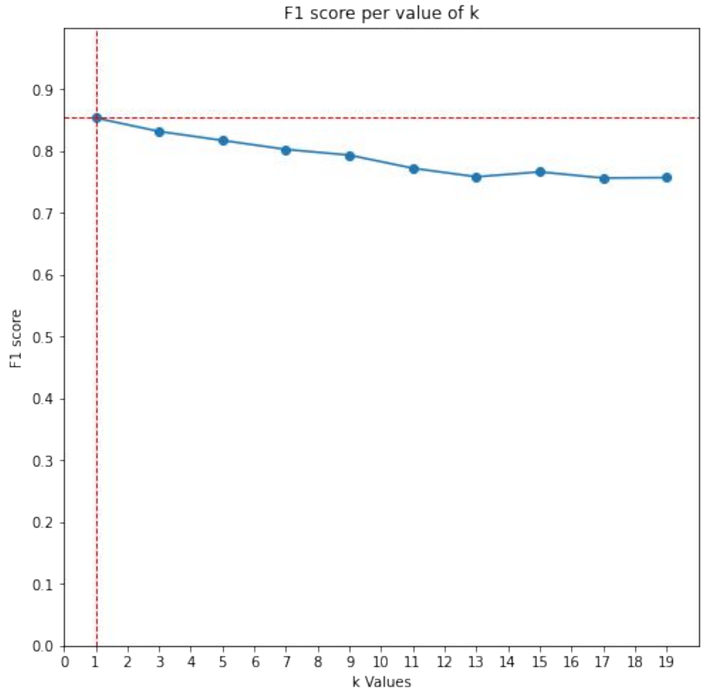
There are 40 soil types in our data set

- We see significant difference in representation between the training and the test dataset for the different soil types

# Models Planned

## Models Planned to Develop

- KNN – k–Nearest Neighbors
- Naive Bayes
- Logistic Regression
- Support Vector Machines
- Decision Tree
- Neural Nets
- Ensemble Models



# Preliminary Results

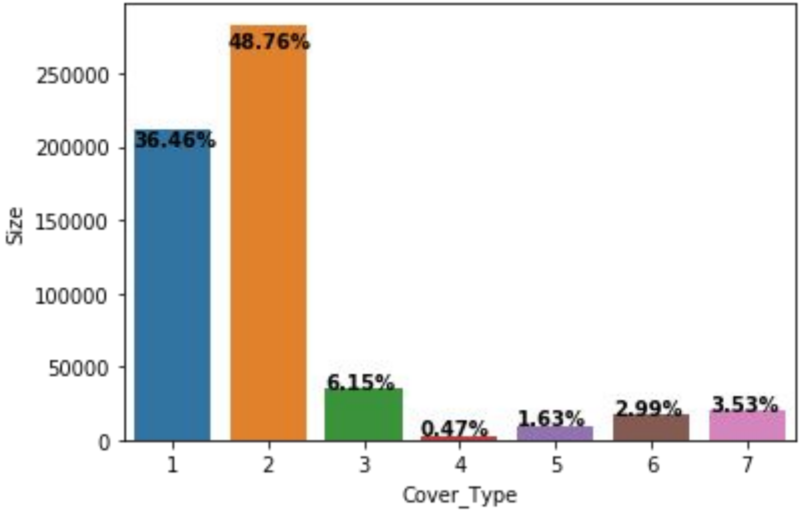
| Model | Description | Training Score | Kaggle Score |
|---|---|---|---|
| **K-Nearest Neighbors** | K = 1 from GridSearch and using Euclidean Distance | 0.854 | **0.710** |
| **Naive Bayes** | Default GaussianNB model | 0.590 | 0.421 |
| **Logistic Regression** | With C = 10 and Penalty = 'L1' from GridSearch | 0.668 | 0.560 |
| **Neural Network** | Default SKLearn model + Early stopping | 0.651 | 0.573 |
| **Decision Tree** | Max Depth of 20 from GridSearch | **0.998** | 0.593 |

# Stuck Points

- Unbalanced class distribution in training data

- Overfitting on training data, lower scores on Kaggle

- Need to refine how data is used by models



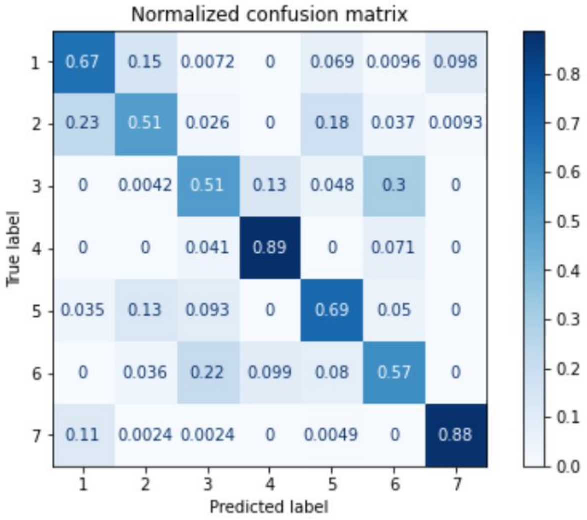
Berkeley SCHOOL OF INFORMATION

# Next Steps

- Address class imbalance
  - Bootstrapping
  - Boosting via AdaBoost

- Address overfitting
  - Tune hyperparameters
  - Ensemble methods e.g. bagging
  - Random seeds for optimizers

- Refine models' use of data:
  - Normalize data
  - Feature engineering



Normalized confusion matrix

| True label \ Predicted label | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0.67 | 0.15 | 0.0072 | 0 | 0.069 | 0.0096 | 0.098 |
| 2 | 0.23 | 0.51 | 0.026 | 0 | 0.18 | 0.037 | 0.0093 |
| 3 | 0 | 0.0042 | 0.51 | 0.13 | 0.048 | 0.3 | 0 |
| 4 | 0 | 0 | 0.041 | 0.89 | 0 | 0.071 | 0 |
| 5 | 0.035 | 0.13 | 0.093 | 0 | 0.69 | 0.05 | 0 |
| 6 | 0 | 0.036 | 0.22 | 0.099 | 0.08 | 0.57 | 0 |
| 7 | 0.11 | 0.0024 | 0.0024 | 0 | 0.0049 | 0 | 0.88 |

```
Overall accuracy: 66.8%
```

# Questions?