Forest Cover

By: Sebastian Banasik and Alexander Jermyn

Introduction

Problem: Identify forest cover based on geological characteristics

Background:

- Forest cover has a large impact on the climate
- Roosevelt National Forest, (Northern Colorado)

Dataset

- Covertype (UC Irvine Machine Learning)
- Identify 7 types of forest cover
  - Type 0: Spruce/Fir
  - Type 1: Lodgepole Pine
  - Type 2: Ponderosa Pine
  - Type 3: Cottonwood/Willow
  - Type 4: Spruce/Fir and Aspen
  - Type 5: Douglas-Fir
  - Type 6: Krummholz

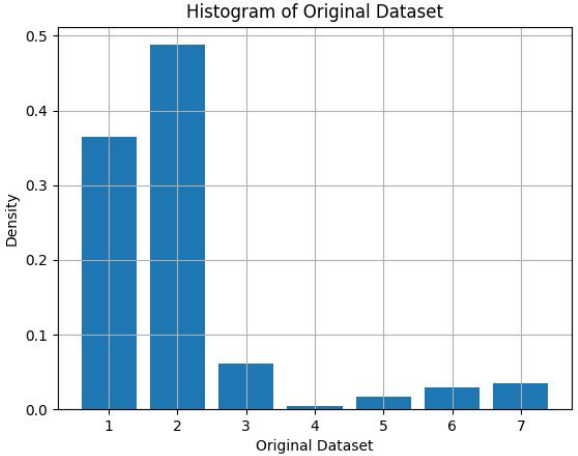
Methodology

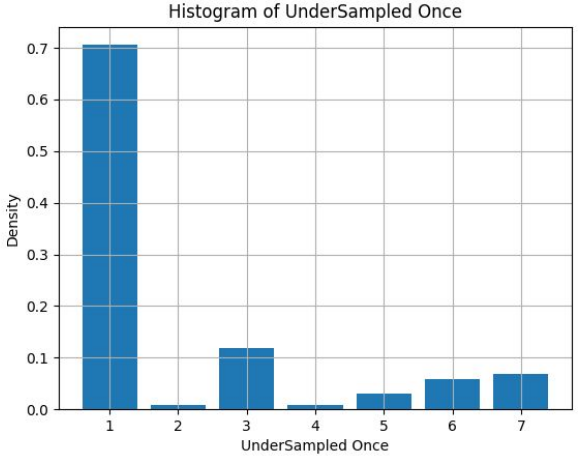
- Exploration
- 4 models
- Evaluation Metrics:
  - Accuracy
  - Recall
  - F1 Score
  - Matthew's Correlation Coefficient

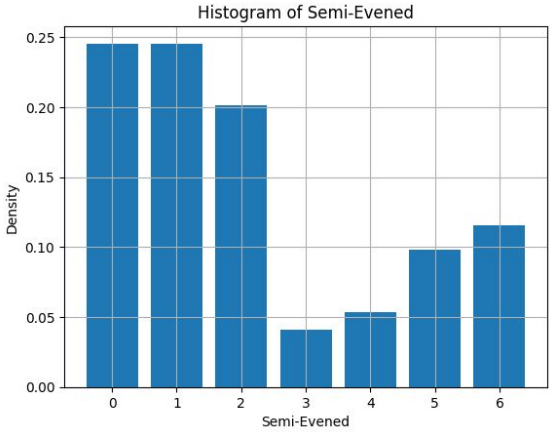
Preprocessing

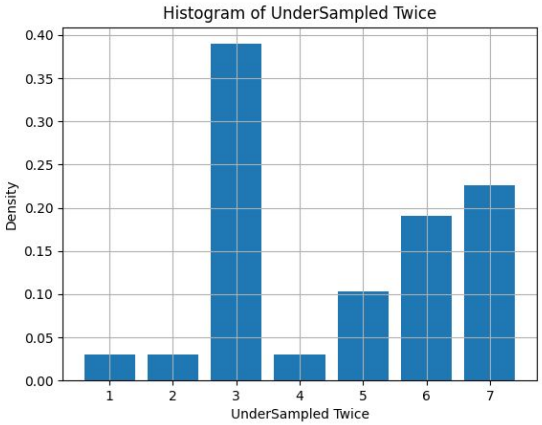
- Dataset included features already 1 hot encoded
- No missing or unlabeled data
- Standard scaling on all features (including 1 hot encoded)
- Created different sampled dataset:
  - No changes
  - Under sampled the largest class
  - Under sampled the 2 largest class
  - Under sampled to even out all classes

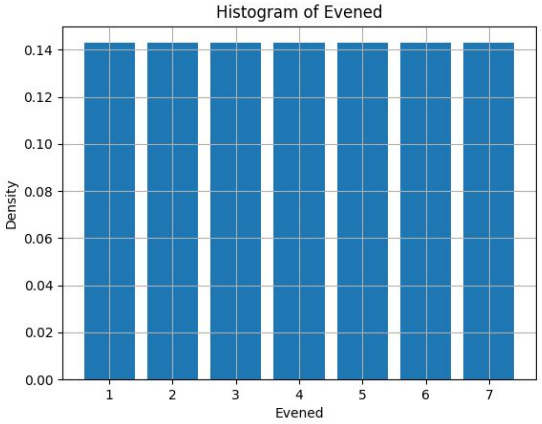
Datasets


Histogram of Original Dataset
Density
Original Dataset


Histogram of UnderSampled Once
Density
UnderSampled Once

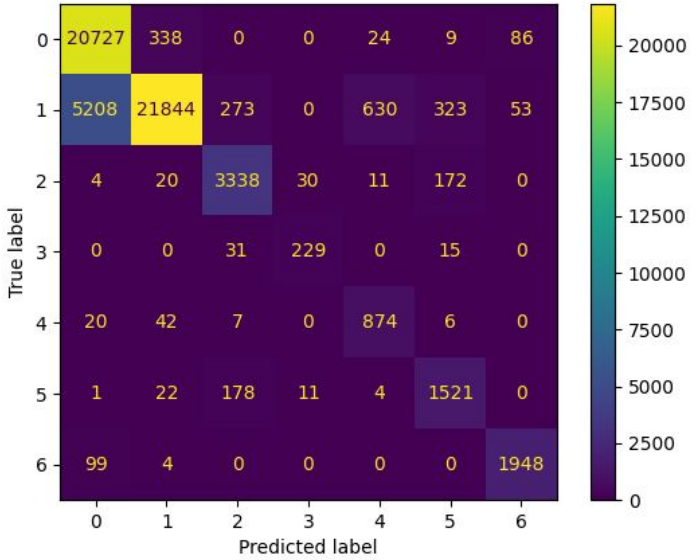

Histogram of Semi-Evened
Density
Semi-Evened


Histogram of UnderSampled Twice
Density
UnderSampled Twice


Histogram of Evened
Density
Evened

K Nearest Neighbors Model

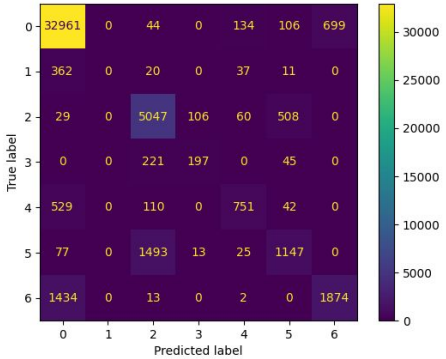
- KNeighborsClassifier
  - Sklearn.neighbors
- Best Parameters:
  - K = 1, Cosine metric
- MCC: 0.807



Linear Models

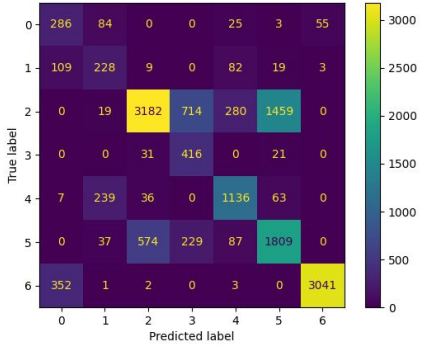
- LogisticRegession, SGDClassifier
  - Sklearn.linear_model
- LogisticRegression Parameters:
  - Penalty: L2, L1, Elasticnet
  - C: 1, .75, .5, .25, 0
  - Max Iteration: 1000, 2000
  - Class Weights: Balanced, Unbalanced
- SGDClassifier Parameters:
  - Penalty: L2, L1, Elasticnet
  - Max Iterations: 1000, 2000
  - Class Weights: Balanced, Unbalanced

Linear Models Evaluation



- Best Model By Score:
  - LogisticRegression:
    - Dataset: Under Sampled Once
    - C: 0.75, Max Iterations: 1000, Penalty: L2, Class Weights: Unbalanced
    - MCC Score: 0.728

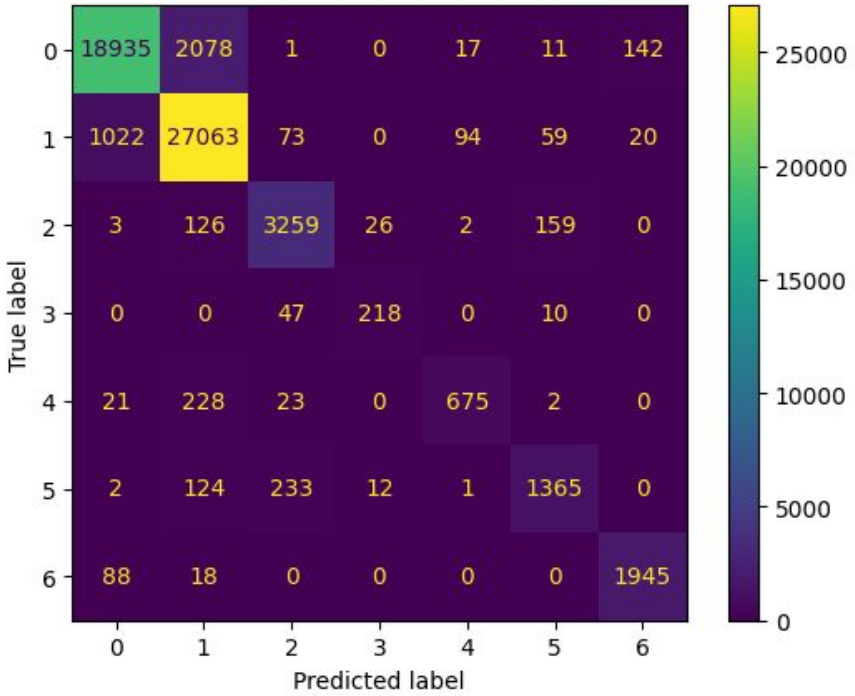
Linear Models Evaluation



- Best Model:
  - LogisticRegression:
    - Dataset: Under Sampled Twice
    - C: 1, Max Iterations: 1000, Penalty: L2, Class Weights: Unbalanced
    - MCC Score: 0.700

Neural Network Model

- Sequential
  - Tensorflow.keras
- Multilayer perceptron
  - 2 hidden layers of 256 nodes
- Regularizations:
  - Early Stopping
- MCC Score: 0.866
- 2 Undersampled, 1 oversampled

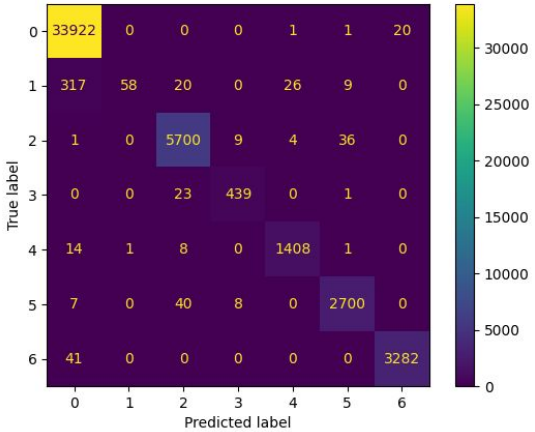


Ensemble Model

- AdaBoostClassifier, with DecisionTreeClassifier
  - Sklearn.ensemble, and sklearn.tree
- AdaBoostClassifier Parameters:
  - Estimators: 50, 100
- DecisionTreeClassifier Parameters:
  - Max Depth: 7, 9, 11, 13, 15, 17
  - Class Weights: Balanced, Unbalanced

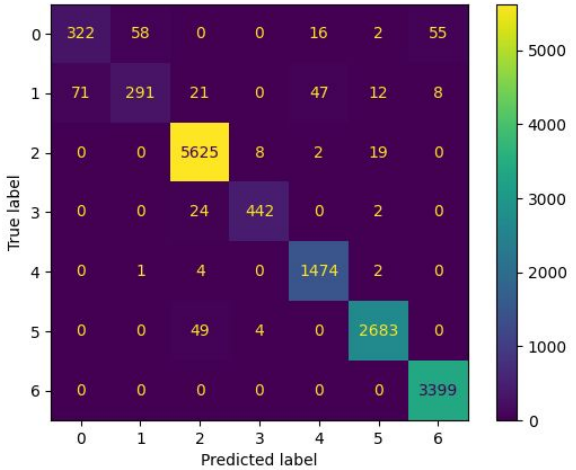
Ensemble Evaluation

- Best Model By Score
  - Dataset: Under Sampled Once
  - Parameters: Max Depth: 15, Estimators: 100, Class Weights: Unbalanced
  - MCC: 0.957



Ensemble Evaluation

- Best Model
  - Dataset: Under Sampled Twice
  - Parameters: Max Depth: 13, Estimators: 100, Class Weights: Unbalanced
  - MCC: 0.929



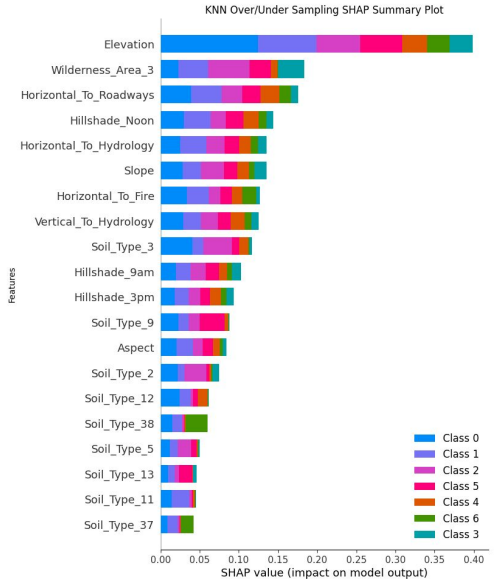
Model Comparisons

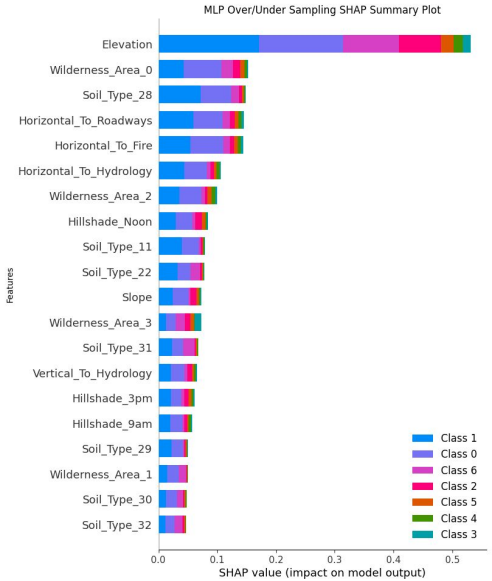
Rankings:

1. Ensemble
   a. MCC Score: 0.929
   b. Dataset: Under Sampled Twice
2. Neural Network
   a. MCC Score: 0866
   b. Dataset: Semi-Evened
3. KNN
   a. MCC Score: 0.807
   b. Dataset: Semi-Evened
4. Linear Models
   a. MCC Score: 0.700
   b. Dataset: Under Sampled Twice

Key Findings

- Balancing class weights didn’t increase performance
- Undersampling of first two classes improves all models
- Elevation is most significant feature





Limitations

- Unequal class representation
- Limited data gathering methods
- Data is outdated
- Data is not well separable

Future Work

- Data from across the globe
- Satellite images/data
- K Medoids Clustering

Acknowledgement

- scikit-learn: machine learning in Python — scikit-learn 1.7.2 documentation
- imbalanced-learn documentation — Version 0.14.0
- TensorFlow
- Covertype - UCI Machine Learning Repository