Forest Cover Type Prediction

Kaggle Competition

Joe Sieber

Table of Contents

Abstract 3

Definition 4

- Introduction 4
- Problem Statement 4
- Metrics 4

Analysis 5

- Data Exploration 5
- Data Visualization 5
- Algorithms and Techniques 5
- Benchmark 5

Methodology 6

- Preprocessing 6
- Implementation 6
- Refinement 6

Results 7

- Model Evaluation and Validation 7
- Justification 7

Conclusion 8

- Critical Features 8
- Reflection 8
- Improvement 8

References 9

Abstract

Definition

Introduction

There are many national and state parks in Colorado. These parks contain some types of forest cover that can be narrowed down to five different groups. The parks can be very remote and hard to navigate the entire park. With such a large area to cover, it would be desireable to be able to know what types of forest cover are a specific locations without having to send people to inspect. This would be very expensive and time consuming. Knowing what type of forest cover is at a given location helps the United States Forest Service (USFS) allocate resources and time to areas that might need more assistance. This also helps the USFS determine if manmade features affect the outcome of what type of forest cover is present for a given location.

The data for this problem was put together by the Colorado State University in order to facilitate answering these questions. This problem was also put forth to the Kaggle community in order to help refine the answer and make better predictions.

Problem Statement

If the forest cover type for all of Colorado was known, it would be easier to allocate resources and time to areas in need. However, classifying the forest cover type for all of the land area in Colorado is a daunting task. This would require an enormous budget and time. The amount of money and time needed makes it cost prohibitive to complete this project. The amount of money saved by knowing this information could never be shown to justify this expense.

Solving this problem with machine learning, although not perfect but reasonably close, would allow a small subset of the land to be sampled in order to create a training set. If successful, this would be a tiny fraction of money need compared to do this manually.

Metrics

In order to determine how successful the machine learning algorithm is the dataset will be split up into three groups. The data will be split into a training, cross validation, and test set. The overall performance of the algorithm will be determined on the test set using multi-class classification accuracy. This metric was used by the Kaggle competition. The reason Kaggle used this metric was because it is a straightforward and easy to interpret metric allows for comparison between algorithms.

Analysis

Data Exploration

The data provided for this analysis has a training and test set already split. Table 1 shows some basic statistics of those data sets.

Number of Training Examples	15120
Number of Test Problems	565892
Number of Features	54
Number of Classifications	7

Table 1 - Basic Statistics

The 54 features are shown in Table 2.

Feature Name	Feature Description
Elevation	Elevation in meters
Aspect	Aspect in degrees azimuth
Slope	Slope in degrees
Horizontal Distance To Hydrology	Horizontal distance to nearest water feature
Vertical Distance To Hydrology	Vertical distance to nearest water feature
Horizontal Distance to Roadways	Horizontal distance to nearest roadway
Hillshade 9 am	Hillshade index at 9 am on summer solstice
Hillshade Noon	Hillshade index at noon on summer solstice
Hillshade 3 pm	Hillshade index at 3 pm on summer solstice
Horizontal Distance to Fire Point	Horizontal distance to nearest ignition point
Wilderness Area	Wilderness area designation (1-4)
Soil Type	Soil type designation (1-40)

Table 2 - Data Set Features

The data does not have any missing information and each example is complete. The 7 cover types represent various types of trees. Table 3 shows those tree types and what the cover type number is. For the training set there are an equal number of examples of each cover type.

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

Table 3 - Cover Types

Data Visualization

The entire training data set was looked at. There were several features that seemed to show interesting findings. However, some of the features seemed to have no influence on the cover type. For example, Aspect seemed to be equally distributed amongst all of the cover types. This is show in Figure 1.

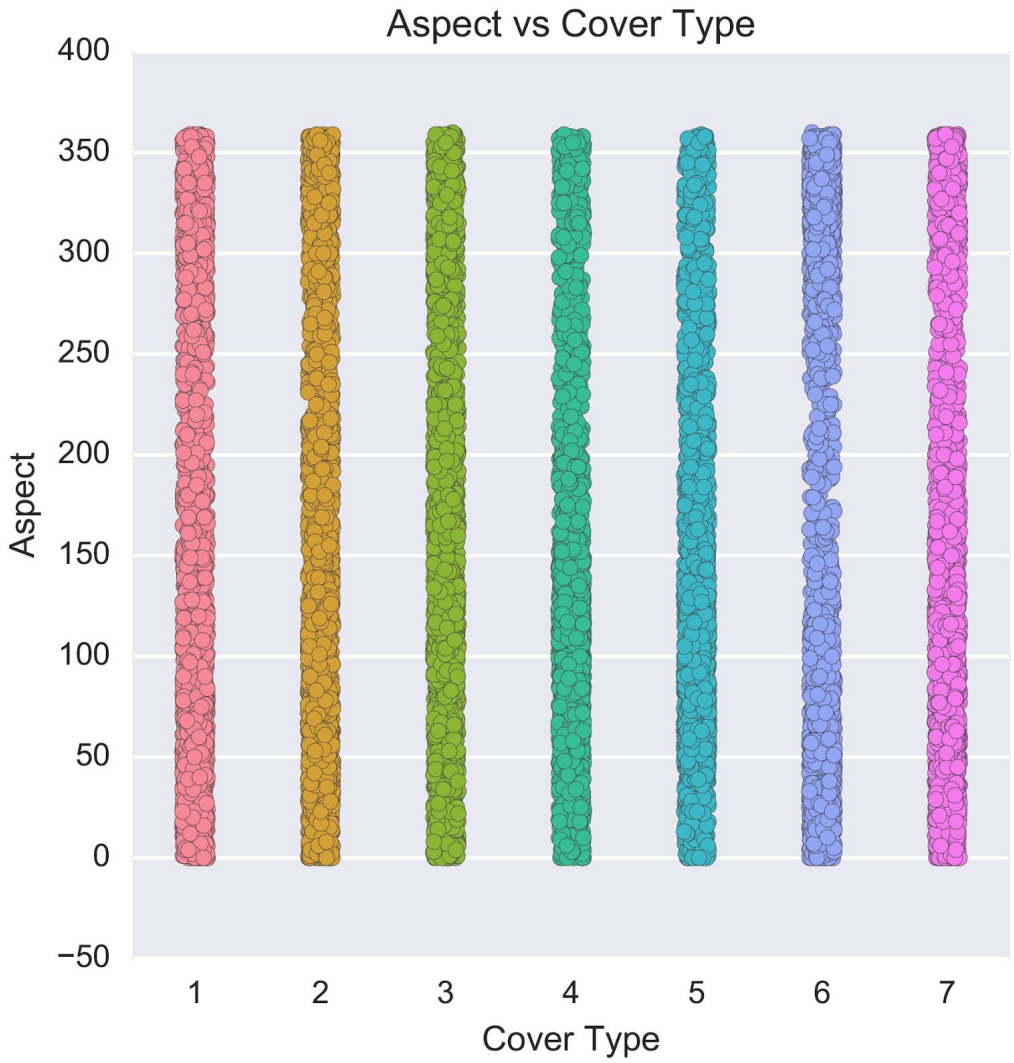

Figure 1 - Aspect vs Cover Type

Algorithms and Techniques

Benchmark

Methodology

Preprocessing

Implementation

Refinement

Results

Model Evaluation and Validation

Justification

Conclusion

Critical Features

Reflection

Improvement

References

https://www.kaggle.com/c/forest-cover-type-prediction