Statement of Work

AI ALGORITHM I

Lin

12/1/2020

Forest Cover Type Prediction

Rationale Statement

Understanding forest composition is a valuable aspect of managing the health and vitality of our wilderness areas. Classifying cover type can help further research regarding forest fire susceptibility and de/reforestation concerns. Forest cover type data is often collected by hand or computed using remote sensing techniques, e.g. satellite imagery. Such processes are both time and resource intensive. This project attempts to predict the predominant type of tree in sections of wooded area, given by data elevation, hydrologic, soil, and sunlight , etc.

Problem

We have been given a total of 54 attributes/features, these attributes contain Binary and Quantitative attributes, and we need to predict which Forest Cover-Type is it from the given features.

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices

The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables such as wilderness areas and soil type.

Data Requirements

Project is based on a famous data set in the machine learning community known as Forest Cover Type available for download in the UCI Machine Learning Repository.

A stratified sample from the original data set to apply the workflow and separate test set to generate final predictions is used as part of a competition in Kaggle.

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest cover type. The seven types are:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

Data

The **training set** (**15120** observations) contains both features and the Cover_Type. The test set contains only the features. You must predict the Cover_Type for every row in the **test set** (**565892** observations).

Data Fields include

Elevation - Elevation in meters

Aspect - Aspect in degrees azimuth

Slope - Slope in degrees

Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features

Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features

Horizontal_Distance_To_Roadways - Horz Dist 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 - Horz Dist to nearest wildfire ignition points

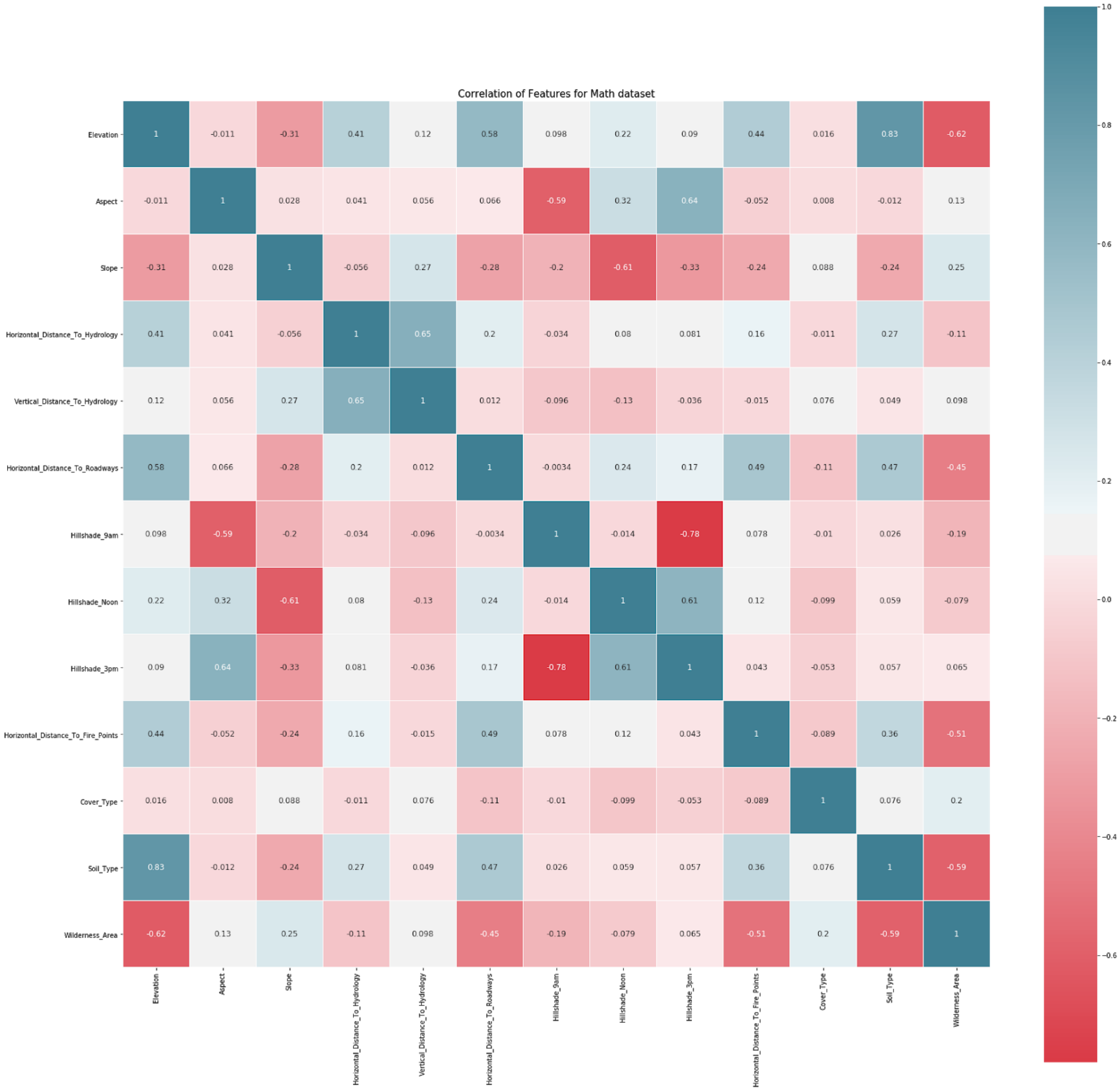
Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation

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

Exploratory Data Analysis

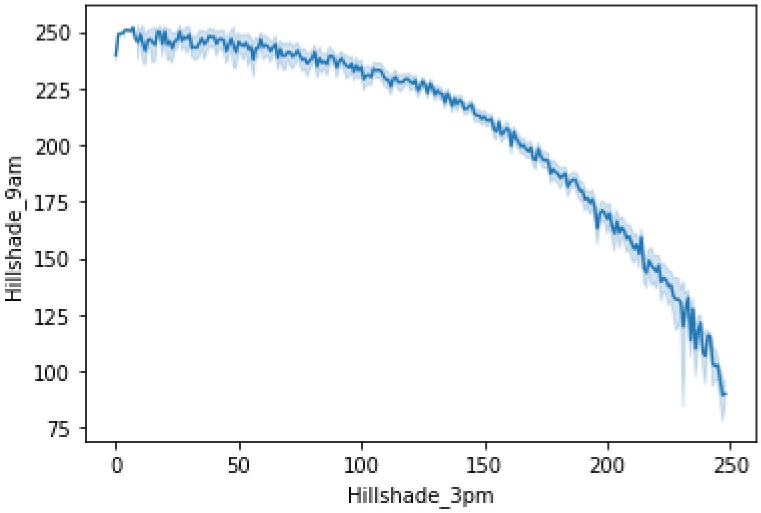
Correlation analysis

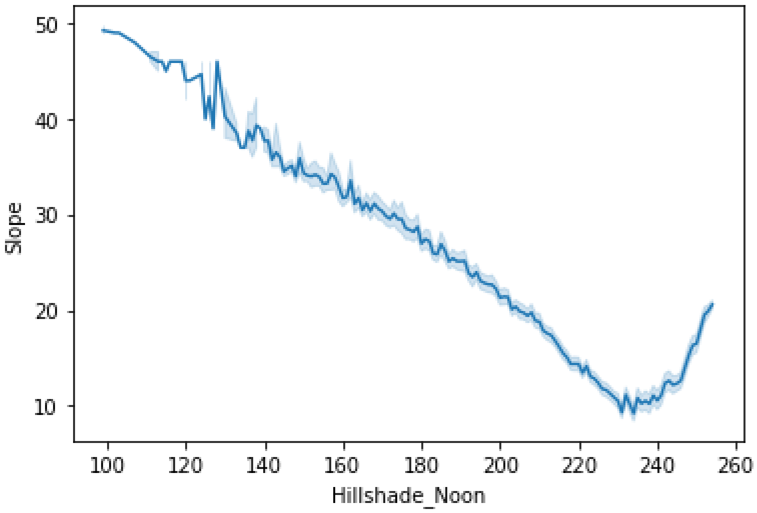


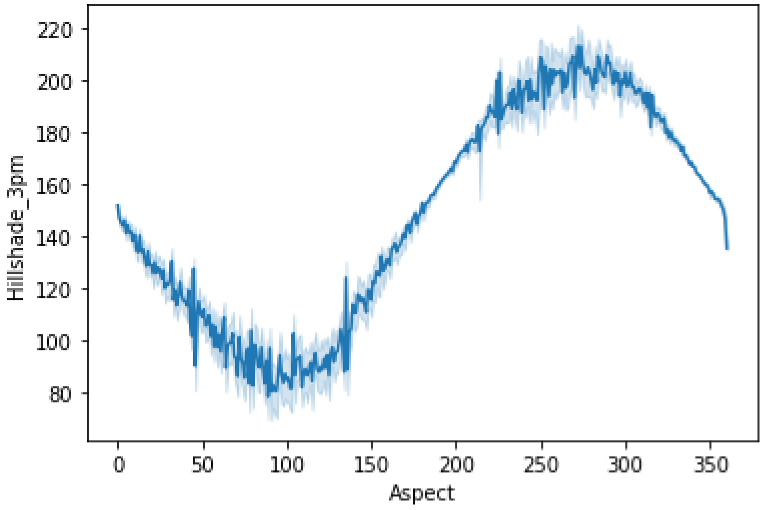
Those features are have high correlations:

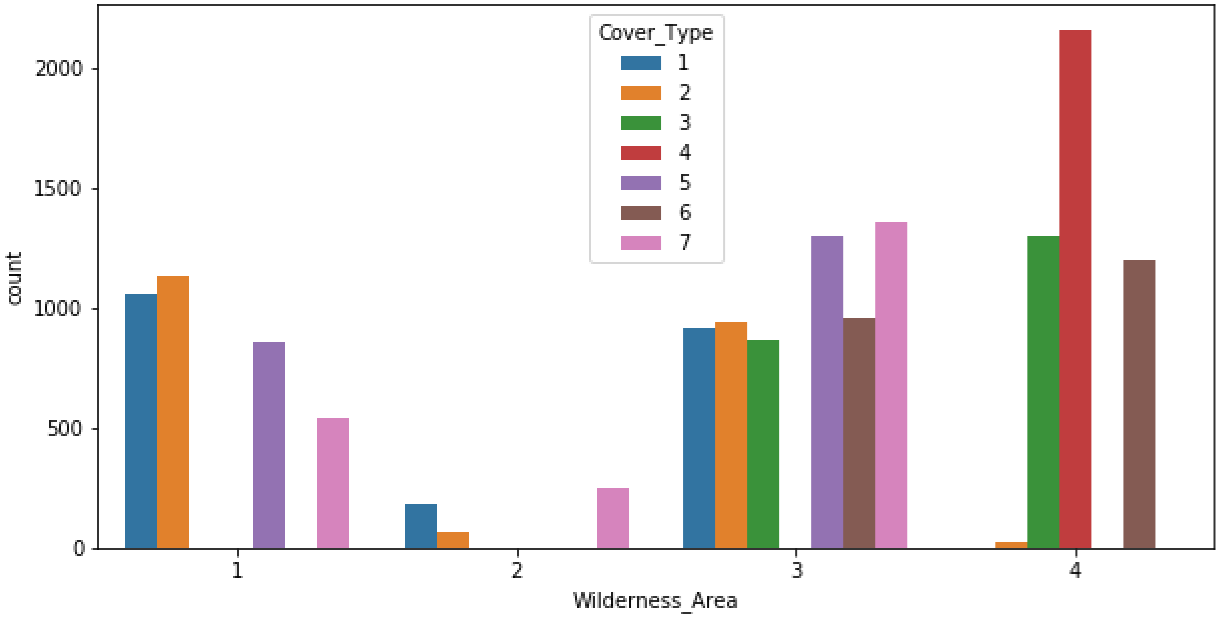
Feature 1	Feature 2	Correlation value
Horizontal_Distance_To_Roadways	Elevation	0.578659
Soil_Type	Elevation	0.826721
Hillshade_3pm	Aspect	0.635022
Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Hydrology	0.652142
Hillshade_3pm	Hillshade_Noon	0.614526

Correlation chart


Hillshade_9am
250
225
200
175
150
125
100
75
0
50
100
150
200
250
Hillshade_3pm


Slope
50
40
30
20
10
100
120
140
160
180
200
220
240
260
Hillshade_Noon





Feature Engineering

Three benefits of performing feature selection before modeling your data are:

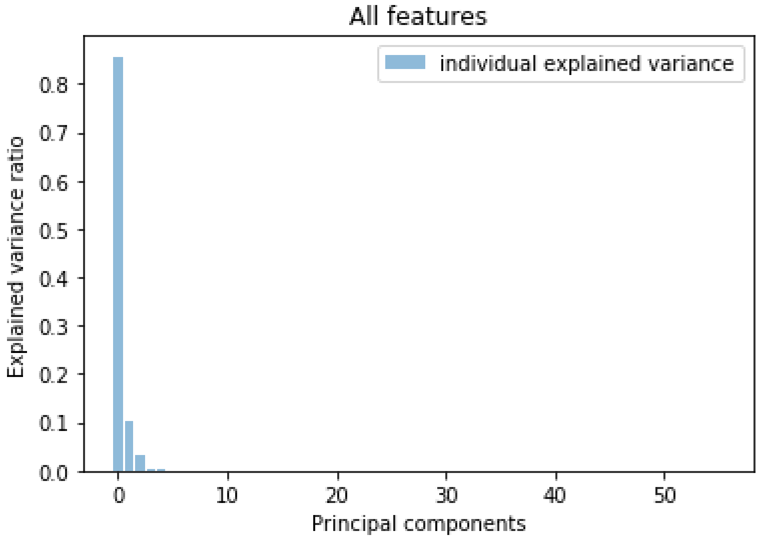
- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.

- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

PCA

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

From below chart, we can know mainly there are 3 independent key features



K Best

SelectKBest will compute the chi2 statistic between each feature of X and y (assumed to be class labels). A small value will mean the feature is independent of y. A large value will mean the feature is non-randomly related to y, and so likely to provide important information. Only k features will be retained.

Below is the sorted features list regarding to their KBest score

Feature	**score**	**p-value**
Soil_Type	2455.599316	0.000000e+00
Wilderness_Area	1875.559841	0.000000e+00
Elevation	1297.239732	4.288061e-277

Horizontal_Distance_To_Roadways	734.718983	1.946781e-155
Horizontal_Distance_To_Fire_Points	395.568520	2.506872e-82
Horizontal_Distance_To_Hydrology	283.960472	2.230028e-58
Slope	134.755529	1.279680e-26
Hillshade_3pm	69.211195	5.933745e-13
Aspect	61.115932	2.669733e-11
Hillshade_9am	34.122835	6.369765e-06
Hillshade_Noon	18.915569	4.308625e-03
Vertical_Distance_To_Hydrology	11.843630	6.554954e-02

Model/Architecture Approach

Methods

The evaluation of the different classification models is going to be based not only on the overall prediction accuracy but also on the accuracy of correctly predicting each of the cover types' classes. To avoid overfitting to the training data, **K-Fold Cross-Validation** technique, which is generally used when estimating how accurate results a method can be assumed to achieve when evaluated on data that is independent of the training data, will be used in this project. The overall accuracy was calculated by taking the average of each correctly predicted type divided by the total number of observations of all five fold cross-validation sets. There are approximately 464,800 observations in the training sets and 116,200 in the test sets.

Algorithms

The most common classification algorithms will be used in this project. The goal is observing predictive accuracy of these algorithms and finding a best one for this dataset. The set of algorithms, which include the k-nearest neighbors algorithm (**k-NN**), Logistic Regression (**LR**), Random Forest (**RF**), and Decision Trees (**DT**).

Feature Selection

Besides the choice of algorithms, an equally important aspect is the dimension of the model. In statistical modeling, it can be translated as finding a simple model with fewer variables that have a high explanatory power. Most data sets contain a notable amount of variables or attributes and it is therefore essential to find a right combination of variables contributing to better describe the overall structure of the data. Feature selection methods will be used to choose such useful variables which are necessary to be included in the model. In this work we will find the right combination of the total 54 attributes to be then used in the model based on EDA, and compare their performances with the respect to prediction accuracy.

Hyperparameter Tuning

In this project, it will also include the hyperparameter tuning on one or two best models and try to improve accuracy of the model.

Evaluation

K-fold Cross validation will be the method used in this project to assess the performance of the algorithms and hyperparameters on the whole dataset. At the end of K-fold cross validation, the average of the performance metric on each of the K iterations substitutes the final performance measure.

As well as the evaluation by cross validation, the prediction results for test dataset will be submitted to the competition in Kaggle to get the kaggle competition score to compare the result with other solutions.