Capstone Project Proposal - Machine Learning Engineer Nanodegree - Forest Cover Type Prediction (https://www.kaggle.com/c/forest-cover-type-prediction)

Domain Background

Creatures on earth depend on forest to live. Forests filter water, create air, provide habitats, etc.. Deforestation is the second leading cause of carbon pollution. In order to protect our forests, many countries designed ‘classification’ programs to keep records of forestlands. The goal of this project is to create a model that can be used to predict and classify forests. A great amount of research was focusing on remote sensing of forest structure. However, our goal in this project is to create a machine learning model that make forest type predictions based on given cartographical variables.

Problem Statement

This project is looking to solve the problem of predicting forest cover type. 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

It is clearly to me this a classification problem and could be solved using the knowledge I learned from Udacity MLND.

Datasets and Inputs

The dataset, both training and testing, could be found on Kaggle.
(https://www.kaggle.com/c/forest-cover-type-prediction/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 are presented below

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

Solution Statement

To solve this problem, I am going to train some classification models from sklearn library. And use the trained the models to predict the forest cover type.

Benchmark Model

While the Kaggle competition has passed its deadline and is no longer active. The competition leader board still accept submission and rank the performance of the algorithms.

Evaluation Metrics

The f1 score was used as an evaluation metric for this problem. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

```
F1 = 2 * (precision * recall) / (precision + recall)
```

Project Design

I am solving this problem in following steps:
- pre-process the data to remove outliers
- fit various models
- tune the models if needed with grid search cross validation techniques
- perform predictions on the test dataset

- ensemble the models predictions to get a final score
- submit the test dataset to the Kaggle competition for benchmarking and evaluation.