**Problem description**

The goal of this project is to predict the forest cover type, which is the predominant kind of tree cover. The study area includes 4 wildness area in the Roosevelt National Forest of northern Colorado, and it is divided into 30 m X 30 m cells. Each cell is given a cover type based on US Forest Service (USFS) Region 2 Resource Information System data. Features for each cell was derived from data obtained from the US Geological Survey and USFS.

**Data**

The training set (15120 observations) contains both features and the label, i.e. Cover\_Type.

There are totally 7 forest cover types, and they are coded into 1 through 7. The seven types are: 1 - Spruce/Fir; 2 - Lodgepole Pine; 3 - Ponderosa Pine; 4 - Cottonwood/Willow; 5 – Aspen; 6 - Douglas-fir; 7 – Krummholz.

There are totally 54 features, which can be grouped into 9 groups. The features are listed and explained as following.

* Elevation is the height of a study point above Earth’s sea level. Higher elevation suggests lower temperature, and temperature is critical for determining tree types.
* Aspect indicates the direction which the hill is facing towards. 0 degree means that hill is facing north, 90 degree east, 180 south and 270 west. Aspect mainly determines the amount of sunlight which the detection point can accept. Hills facing south will accept more sunlight than those facing north.
* Slope is the degree of the slope where detection point locates.
* Horizontal\_Distance\_To\_Hydrology and Vertical\_Distance\_To\_Hydrology are horizontal or vertical distance to nearest surface water features, respectively.
* Horizontal\_Distance\_To\_Roadways is the horizontal distance to nearest roadway, and this parameter accounts for human activity.
* Hillshade\_9am, Hillshade\_Noon, and Hillshade\_3pm are hillshade indices at respective time points. Actual data is in 0 to 255 scale with 255 meaning complete brightness and 0 meaning total darkness.
* Horizontal\_Distance\_To\_Fire\_Points is the horizontal distance to nearest wildfire ignition points. This variable accounts for the dryness of detection point. Drier area is more favorable for wildfire ignition.
* Wilderness\_Area indicates the location of wilderness area, including Rawah Wilderness Area, Neota Wilderness Area, Comanche Peak Wilderness Area and Cache la Poudre Wilderness Area. Each wilderness area is encoded into a binary column. 0 means absence and 1 means presence.
* Soild\_Type designates the soil type. There are totally 40 soil types, and each type is encoded into a binary column. 0 means absence and 1 means presence.

**Model**

Decision tree and random forest were used to predict the labels. Matlab function fitctree() was used for building decision tree. 10-fold cross validation was used for validating results. Function TreeBagger() was used for building random forest.

**Results**

Using fitctree() function from matlab, we can build a decision tree based on our data. The decision tree is shown in figure 1.

Figure 1. Decision tree visualization

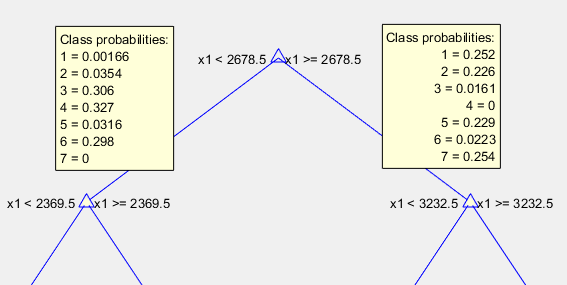


10-fold cross validation gives an error of 0.2110, meaning that ~80% of training data can be classified correctly with this decision tree.

If we zoom in on the decision tree (Figure 2), we can get more details about each node as well as the importance of each classifier. Figure 2 also gives us the distribution probability after classification by the first node. From figure 2, we can see feature x1 (elevation) is the most powerful classifier for this dataset since the first three nodes are all dependent on x1 values.

The first node divides data into two parts based on whether x1 is less or more than 2678.5. Based on the probability, forest type 1, 2, 5 and 7 barely exist in the population x1 < 2678.5. On the contrary, forest type 3, 4 and 6 barely exist in the population x1 >=2678.5.

Figure 2. Zoomed-in view of decision tree showing first three nodes



In order to further analyze which feature is important for classifying forest types, each feature was left out for building a new decision tree and testing feature importance. 10-fold cross-validation was used for calculating validation error. The results are shown in figure 3. In figure 3, x-axis indicates which feature was left out, and the y-axis denotes cross validation error. Note that x-axis only contains 12 features instead of 54 features, because feature Soil\_Type was left out together as a group (#12), and same with Wilderness\_Area (#11). Here higher cross validation error indicates that the feature has more importance.

Through figure 3, we can conclude that elevation is the most important feature for determining forest types. Besides elevation, three other important features are also very important: Horizontal\_Distance\_To\_Roadways, Horizontal\_Distance\_To\_Fire\_points and Soil\_Type. We can visualize the relationship between these four features and labels in figure 4.

Figure 3. Feature Importance

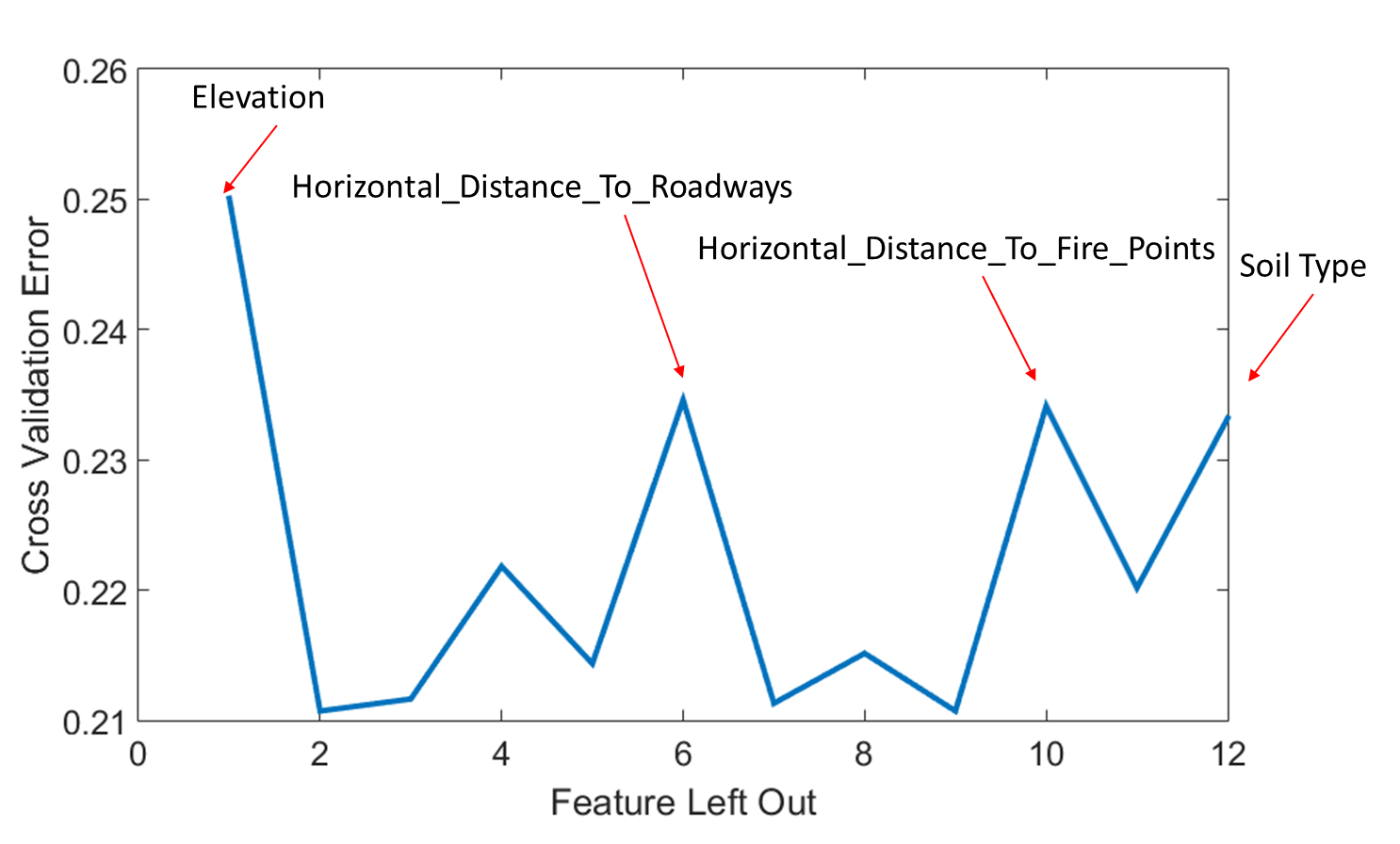
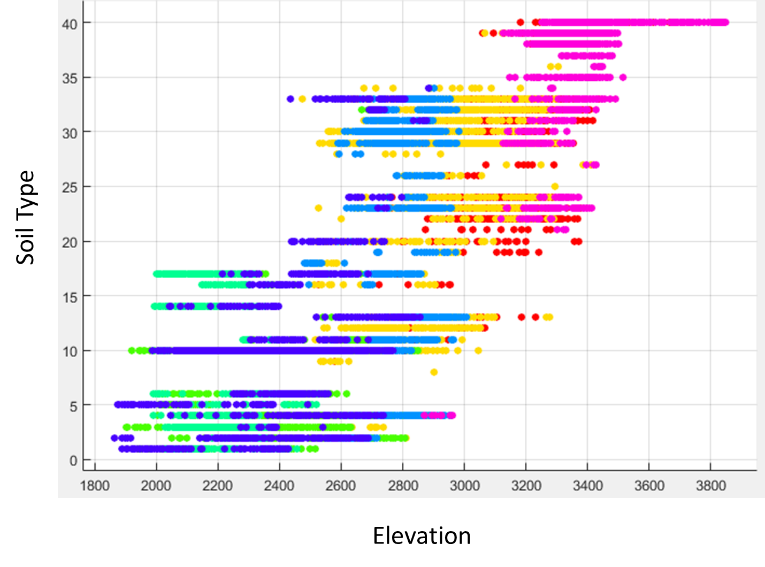
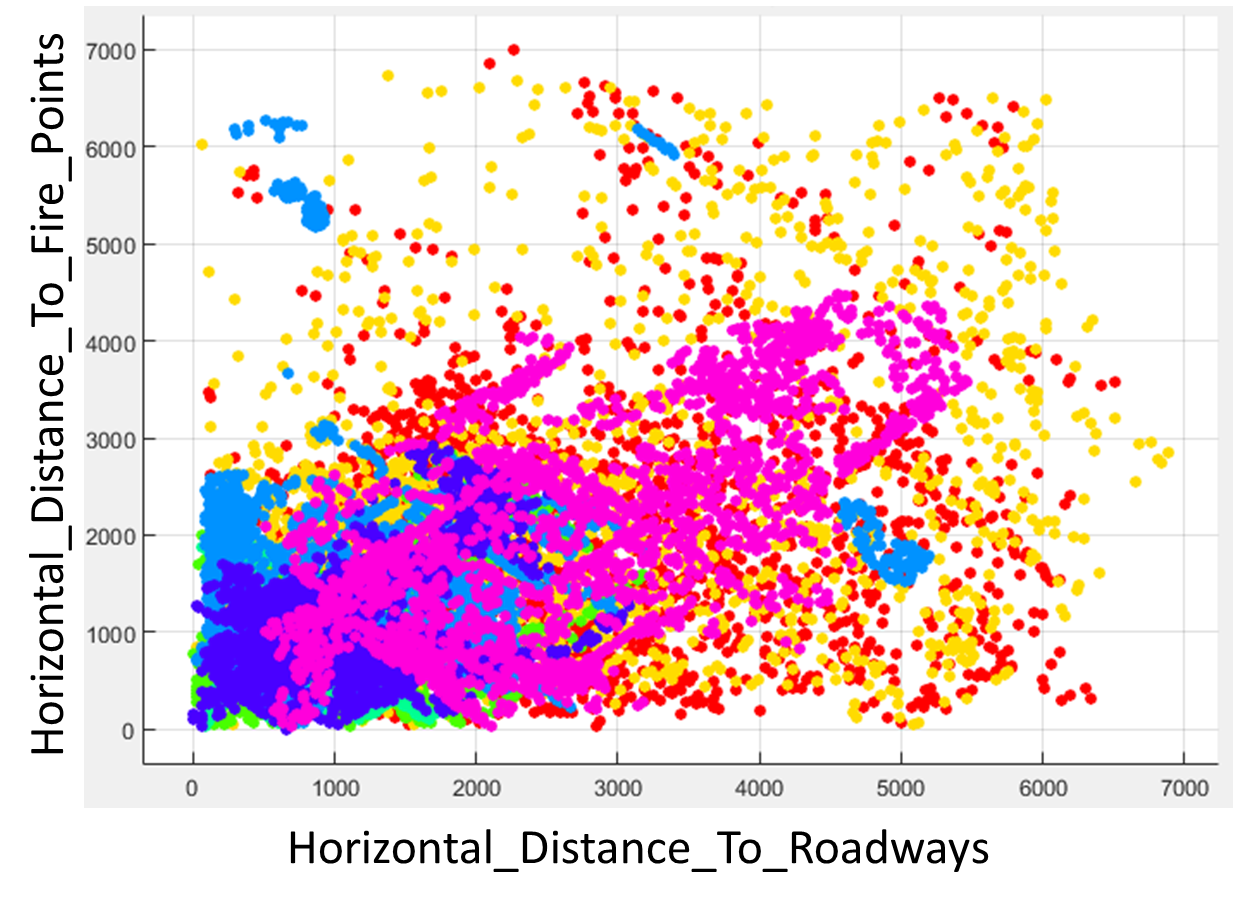


Figure 4.

Random forest is an ensemble method for classification and usually produces better result than one single decision tree. Here random forest was used to fit the data, and the out-of-bag classification error was compared with cross validation error from single decision tree result (figure 5). Averaging ~10 trees will give an out-of-bag classification error lower than one decision tree. Averaging 100 trees result in an out-of-bag classification error of ~0.15.

Figure 5. Out-of-bag classification error with number of grown trees



**Discussion**

This problem is a classification problem. Linear methods such as linear discriminant analysis or logistic regression are not suitable for this particular problem because of the uniqueness of the features. There are 44 features (4 wilderness areas and 40 soil types) that are binary data, only consisting 0 or 1. Therefore they are not in a Gaussian distribution. Decision tree does not assume any distribution, so it is more suitable for this large set of binary data.

Decision tree gives a 10-fold cross validation error of 0.21, meaning that ~80% data achieved the right classification labels. This is a relatively good result, considering that randomly assigning 7 labels will give an extremely high error rate. Part of the reasons for this accuracy is the relatively large training data with 15120 data entries but only 54 features. In fact, some of the soil type features do not have any positive values, such as Soil\_Type7 and Soil\_Type15, and this further reduces the dimension.

Feature importance analysis revealed several important classifiers in this problem. The most important feature is elevation. This is reasonable because elevation determines the temperature where tree grows: the temperature decreases as elevation increases. The leaves of trees will be in a spread shape in area with higher temperature, but needle-like shapes in areas with lower temperature. Based on the decision tree result, elevation of 2678.5 meters separates forests into two major groups. Another three important features are soil type, horizontal distance to roadways and horizontal distance to fire points. Soil type is correlated with forest cover type mainly because of two reasons. First the nutrients in soil can determine what kind of trees can grow. Second soil types reflect other environmental factors such as water supply, animal behavior etc., which can also help determine tree type. Horizontal distance to fire points reflects the dryness of the area, which determines the amount of water the trees can achieve and apparently impacts forest cover type. As for horizontal distance to roadways, this parameter may not actually determine the forest cover type, but only has a correlation with forest cover type. It is possible that roadways are built in particularly chosen sites. In conclusion, forest cover type seems to be most closely related with temperature, soil and humidity.

Decision tree method has some flaws such as overfitting and instability. Random forest becomes a better method because of its ensemble nature. It has many advantages over decision tree. Here using random forest as our model resulted in smaller error and gave a better prediction than decision tree. In addition, the error rate decreased quickly with the number of grown trees, and averaging ~10 trees already gave better results than decision tree. Therefore it does not cost too much more than decision tree computationally.