Forest cover type prediction

In this competition you are asked to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (as opposed to remotely sensed 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. Training set (15120 observations), where indicationg cover type, and test set(56589 observations), without indicating cover type.

Data Fields

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 (see Kaggle for a qualitative description of each soil type)

Cover_Type (7 types, integers 1 to 7) - cover type

Source

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

First, we will work with the training set, in which we have . Let's display the first six rows, in order to see how to construct data.

```
##   Id Elevation Aspect Slope Horizontal_Distance_To_Hydrology
## 1  1      2596     51     3                              258
## 2  2      2590     56     2                              212
## 3  3      2804    139     9                              268
## 4  4      2785    155    18                              242
## 5  5      2595     45     2                              153
## 6  6      2579    132     6                              300
##   Vertical_Distance_To_Hydrology Horizontal_Distance_To_Roadways
## 1                              0                             510
## 2                             -6                             390
## 3                             65                            3180
## 4                            118                            3090
## 5                             -1                             391
## 6                            -15                              67
##   Hillshade_9am Hillshade_Noon Hillshade_3pm
## 1           221            232           148
## 2           220            235           151
## 3           234            238           135
## 4           238            238           122
## 5           220            234           150
## 6           230            237           140
##   Horizontal_Distance_To_Fire_Points Wilderness_Area1 Wilderness_Area2
## 1                               6279                1                0
## 2                               6225                1                0
## 3                               6121                1                0
## 4                               6211                1                0
## 5                               6172                1                0
## 6                               6031                1                0
##   Wilderness_Area3 Wilderness_Area4 Soil_Type1 Soil_Type2 Soil_Type3
## 1                0                0          0          0          0
## 2                0                0          0          0          0
## 3                0                0          0          0          0
## 4                0                0          0          0          0
## 5                0                0          0          0          0
## 6                0                0          0          0          0
##   Soil_Type4 Soil_Type5 Soil_Type6 Soil_Type7 Soil_Type8 Soil_Type9
```

```
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   Soil_Type10 Soil_Type11 Soil_Type12 Soil_Type13 Soil_Type14 Soil_Type15
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           1           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   Soil_Type16 Soil_Type17 Soil_Type18 Soil_Type19 Soil_Type20 Soil_Type21
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   Soil_Type22 Soil_Type23 Soil_Type24 Soil_Type25 Soil_Type26 Soil_Type27
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   Soil_Type28 Soil_Type29 Soil_Type30 Soil_Type31 Soil_Type32 Soil_Type33
## 1           0           1           0           0           0           0
## 2           0           1           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           1           0           0           0
## 5           0           1           0           0           0           0
## 6           0           1           0           0           0           0
##   Soil_Type34 Soil_Type35 Soil_Type36 Soil_Type37 Soil_Type38 Soil_Type39
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   Soil_Type40 Cover_Type
## 1           0          5
## 2           0          5
## 3           0          2
## 4           0          2
## 5           0          5
## 6           0          2
```

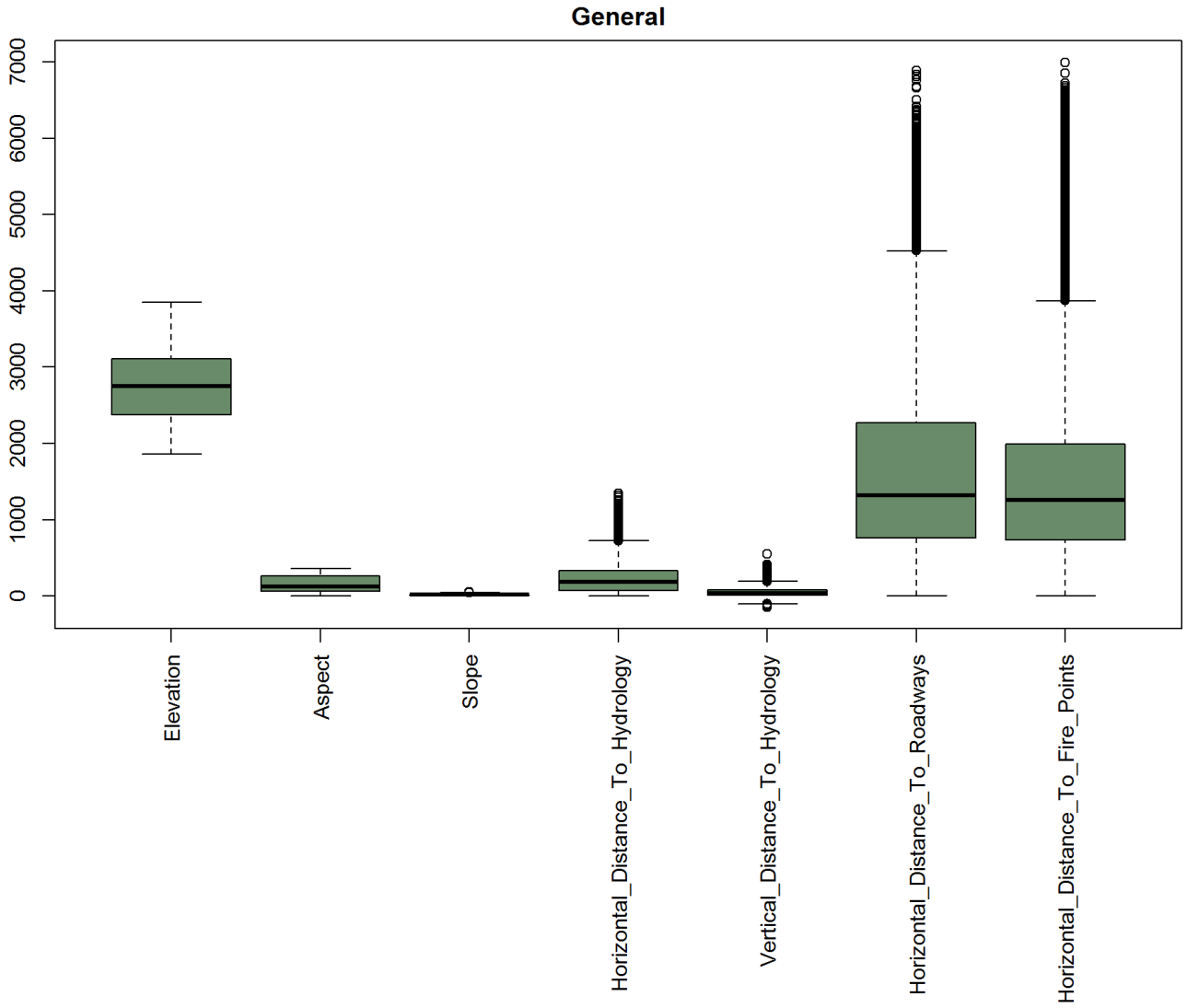
We see unwieldy amount of data is mainly due to the variable Soil_Type, which in this case takes the form of so-called «Dummy variable». There are 40 soil types and therefore we can see 40 columns. If we want define what is the soil type for particular observation, we must to find “1” in one of forty “Soil_Types” columns. The situation is similar for the “Wilderness_Area” variable. So, transform variables “Soil_Type” and “Wilderness_Area” to a more compact form.

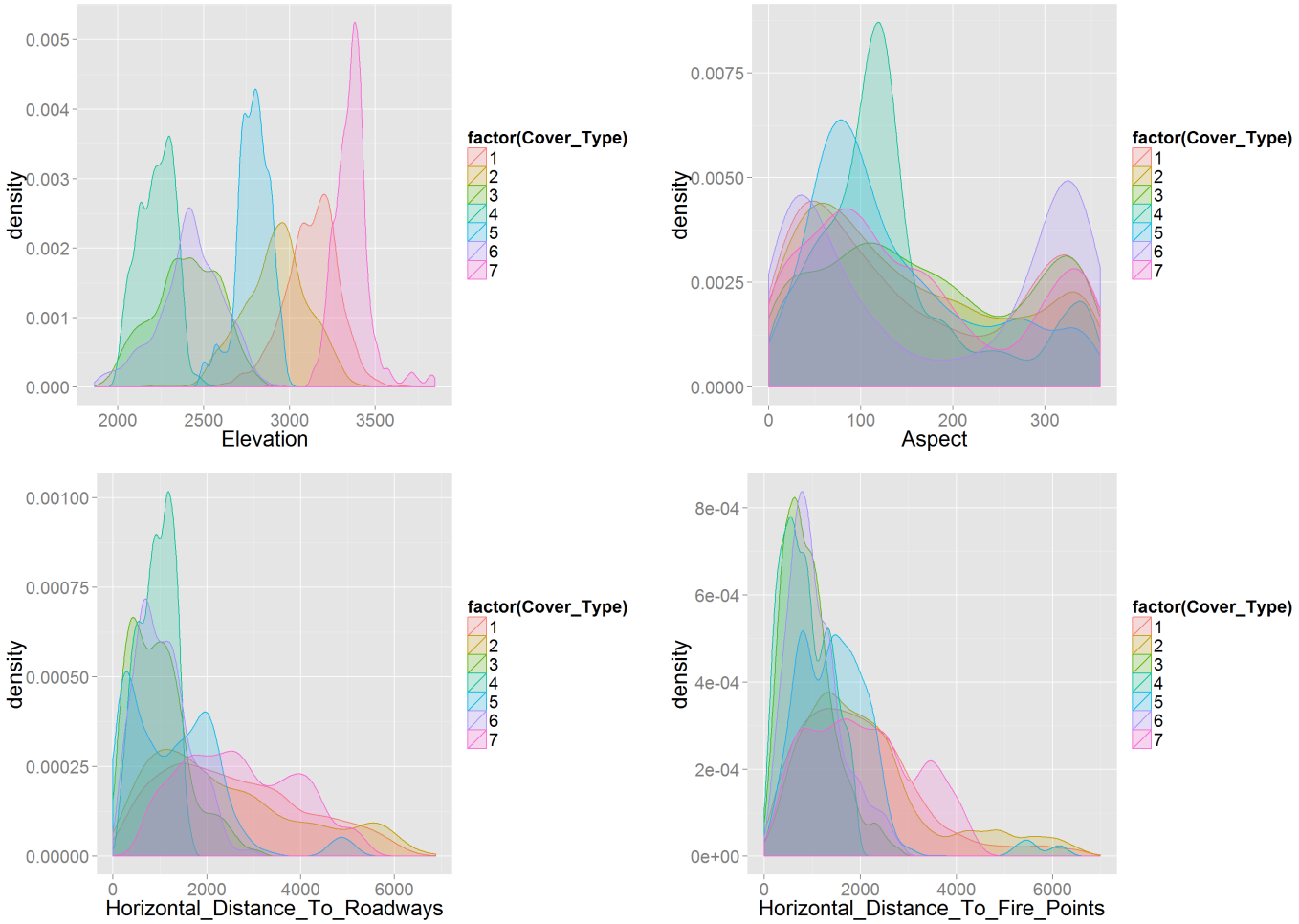
```
##   Elevation Aspect Slope Horizontal_Distance_To_Hydrology
## 1      2596     51     3                              258
## 2      2590     56     2                              212
## 3      2804    139     9                              268
## 4      2785    155    18                              242
## 5      2595     45     2                              153
## 6      2579    132     6                              300
##   Vertical_Distance_To_Hydrology Horizontal_Distance_To_Roadways
## 1                              0                             510
## 2                             -6                             390
## 3                             65                            3180
## 4                            118                            3090
## 5                             -1                             391
## 6                            -15                              67
##   Hillshade_9am Hillshade_Noon Hillshade_3pm
## 1           221            232           148
## 2           220            235           151
## 3           234            238           135
## 4           238            238           122
## 5           220            234           150
## 6           230            237           140
```

```
##   Horizontal_Distance_To_Fire_Points Soil_Type Wilderness_Area Cover_Type
## 1                               6279        27               1          5
## 2                               6225        27               1          5
## 3                               6121        11               1          2
## 4                               6211        28               1          2
## 5                               6172        27               1          5
## 6                               6031        27               1          2
```

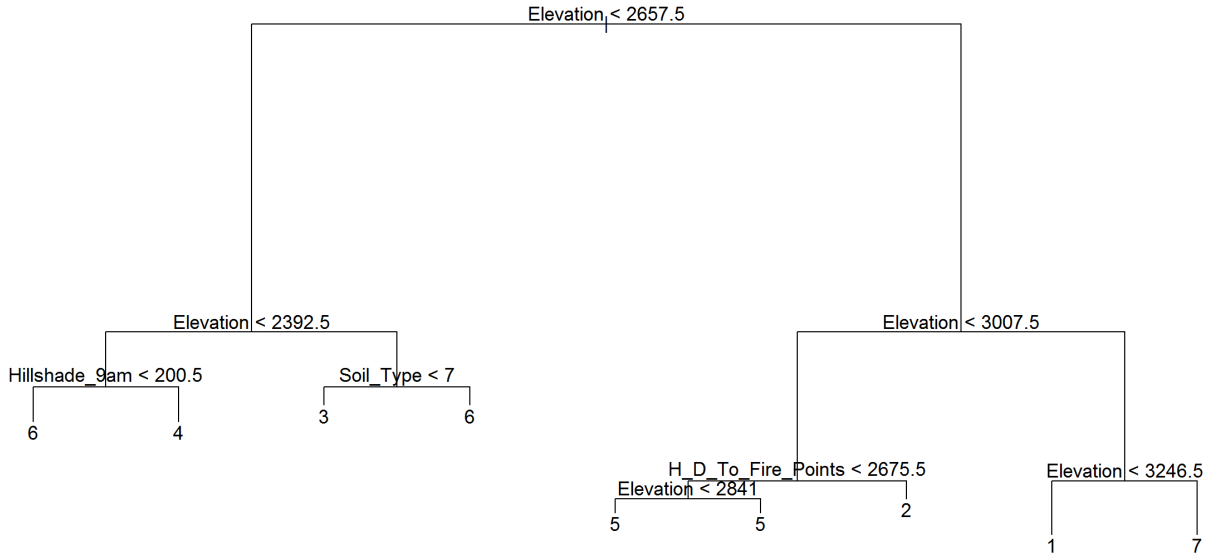
After transforming we recieve more compact form of these variables. Now we have one column for each variables “Soil_Type” and “Wildness_area”, instead of 40 columns that we had in previous case.

Let’s use a plots to see the structure of data.





It seems like elevation will be a pretty good variable to use for classification. Unfortunately, all of the variables are not this good. If build a decision tree, we will see that it ignores most of the variables. The decision tree algorithm basically works by choosing variables which best split the data and creating 'rules' for classification. T



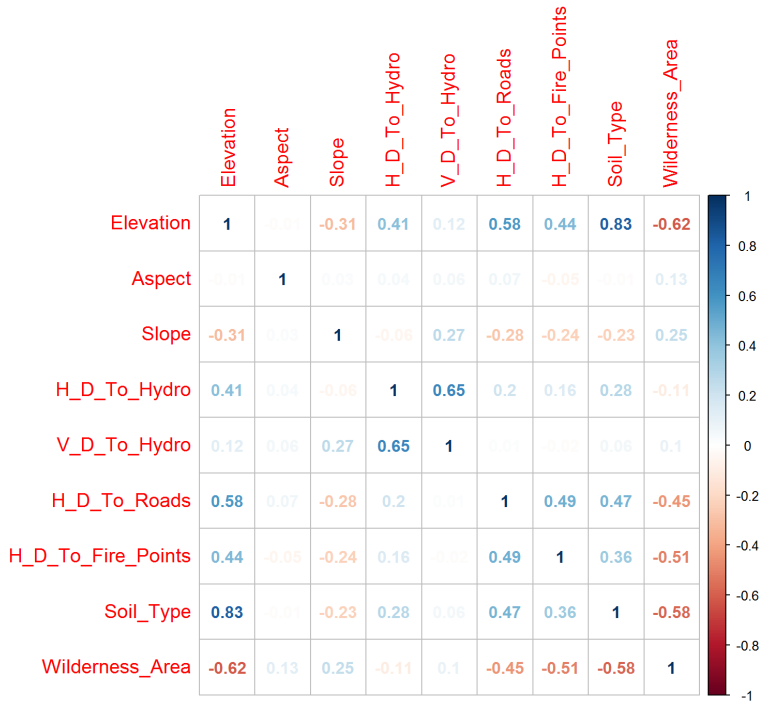
It basically decided that elevation was a great way to split up the data. If the elevation is less than 2657,5 the classes are most likely to be 3, 4, 6. If the elevation is greater than 2657,5 the classes are most likely to be 5, 2, 1 or 7. If use this tree to predict new data it is only correct about 60% of the time. So it didn't do so great as a classifier but it could still be useful to make educated guesses.

```
##           Reference
## Prediction  1  2  3  4  5  6  7
##          1 50  2  0  0 15  2 17
##          2 26 10  0  0 30  7  7
```

```
##          3   0   0 36 17   5 50   0
##          4   0   0  3 83   0 11   0
##          5   0   0  2  0 83  6   0
##          6   0   0 13 16   7 55   0
##          7 16   0   0  0   0  0 91
```

```
##  Accuracy
## 0.6181818
```

Let's build a correlation matrix.



Split the train set into training (70%) and testing (30%) groups. This is very useful because, it allows to check how model works and check its accuracy, and then apply model to the test set.

Let's use Random Forest algoritm for prediction. Random forests, as the name suggests, is a group of decision trees (a forest). It basically works by creating a bunch of decision trees by randomly selecting which variables to use then making predictions by outputting the mode of the class from those individual trees. As we worked with file train, now we can compere the result of prediction with original classification in file train.

```
##    Cover_Type Prediction
## 1           5          5
## 2           2          1
## 3           2          2
## 4           1          1
## 5           1          1
## 6           2          2
## 7           2          2
## 8           1          1
## 9           2          2
## 10          2          2
## 11          2          2
## 12          2          2
## 13          2          2
## 14          2          2
## 15          2          2
```

Calcultion accuracy of our model.

```
##           Reference
## Prediction   1   2   3   4   5   6   7
##          1  60  29   0   0   0   0   7
##          2  13  38   0   0   7   0   0
##          3   0   3  76   3   1  20   0
##          4   0   0  12  92   0   7   0
##          5   2   8   1   0  82   3   0
##          6   1   1  19   2   1  61   0
##          7  10   1   0   0   0   0 100
```

```
##  Accuracy
## 0.7712121
```

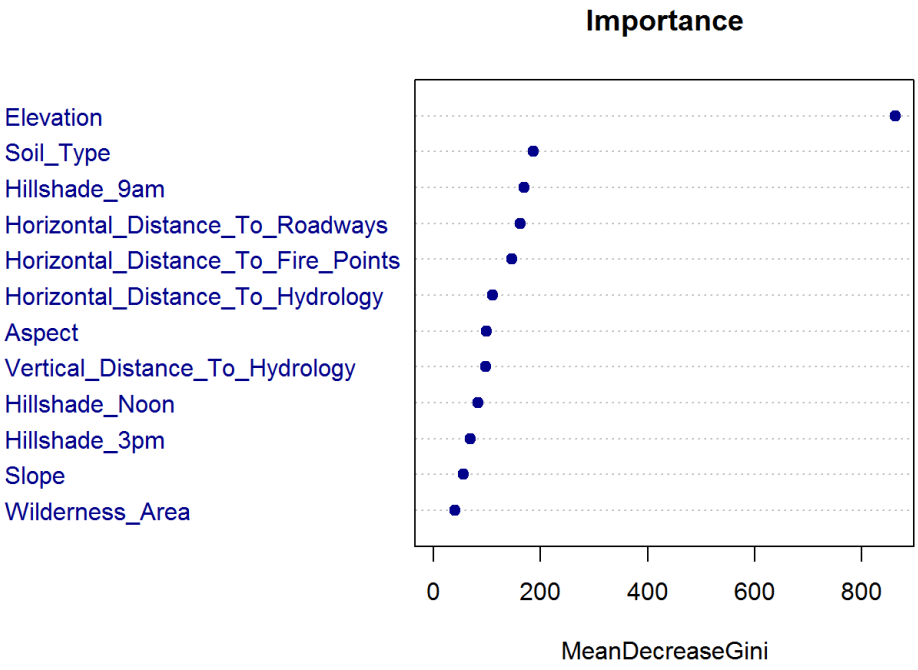
After testing our model we can move on to the next step. Let's apply this model to file test. In previous case we divide train test into training(70%) and testing(30%), here instead training test we will use whole file train , instead testing we will use whole file test. Test file is identical to file train, instead in file test not specified cover type. We will make the same operations with file test as we made before with file train. Transform variables "Soil Type"" and "Wilderness Area".

```
##   Elevation Aspect Slope Horizontal_Distance_To_Hydrology
## 1      2680    354    14                                0
## 2      2683      0    13                                0
## 3      2713     16    15                                0
## 4      2709     24    17                                0
## 5      2706     29    19                                0
## 6      2699     21    18                               30
##   Vertical_Distance_To_Hydrology Horizontal_Distance_To_Roadways
## 1                              0                             2684
## 2                              0                             2654
## 3                              0                             2980
## 4                              0                             2950
## 5                              0                             2920
## 6                              3                             2890
##   Hillshade_9am Hillshade_Noon Hillshade_3pm
## 1           196            214           156
## 2           201            216           152
## 3           206            208           137
## 4           208            201           125
## 5           210            195           115
## 6           206            200           127
##   Horizontal_Distance_To_Fire_Points Wilderness_Area Soil_Type
## 1                               6645               1        29
## 2                               6675               1        29
## 3                               6344               1        29
## 4                               6374               1        29
## 5                               6404               1        29
## 6                               6434               1        29
```

Similarly as in the previous case apply the algorithm Random forest. Save the answer in file and look at the first few elements of answer:

```
##       Id Cover_Type
## 1  15121          5
## 2  15122          1
## 3  15123          1
## 4  15124          1
## 5  15125          1
## 6  15126          1
## 7  15127          1
## 8  15128          1
## 9  15129          1
## 10 15130          1
## 11 15131          1
## 12 15132          1
## 13 15133          1
## 14 15134          2
## 15 15135          2
## 16 15136          2
## 17 15137          1
## 18 15138          1
## 19 15139          1
## 20 15140          2
```

The importance of the variables in predicting:


Importance
Elevation
Soil_Type
Hillshade_9am
Horizontal_Distance_To_Roadways
Horizontal_Distance_To_Fire_Points
Horizontal_Distance_To_Hydrology
Aspect
Vertical_Distance_To_Hydrology
Hillshade_Noon
Hillshade_3pm
Slope
Wilderness_Area
0
200
400
600
800
MeanDecreaseGini