forest_cover-type

May 11, 2019

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
train = pd.read_csv("/Users/DoryChen/Downloads/forest-cover-type-prediction/train.csv")
Test=pd.read_csv("/Users/DoryChen/Downloads/forest-cover-type-prediction/test.csv")
test=Test
```

In [5]:
```
train.head(10)
```

Out[5]:
```
   Id  Elevation  Aspect  Slope  Horizontal_Distance_To_Hydrology  \
0   1       2596      51      3                               258
1   2       2590      56      2                               212
2   3       2804     139      9                               268
3   4       2785     155     18                               242
4   5       2595      45      2                               153
5   6       2579     132      6                               300
6   7       2606      45      7                               270
7   8       2605      49      4                               234
8   9       2617      45      9                               240
9  10       2612      59     10                               247

   Vertical_Distance_To_Hydrology  Horizontal_Distance_To_Roadways  \
0                               0                              510
1                              -6                              390
2                              65                             3180
3                             118                             3090
4                              -1                              391
5                             -15                               67
6                               5                              633
7                               7                              573
8                              56                              666
9                              11                              636

   Hillshade_9am  Hillshade_Noon  Hillshade_3pm  \
0            221             232            148
1            220             235            151
```

```
2           234               238             135
3           238               238             122
4           220               234             150
5           230               237             140
6           222               225             138
7           222               230             144
8           223               221             133
9           228               219             124

   Horizontal_Distance_To_Fire_Points  Wilderness_Area1  Wilderness_Area2  \
0                                6279                 1                 0
1                                6225                 1                 0
2                                6121                 1                 0
3                                6211                 1                 0
4                                6172                 1                 0
5                                6031                 1                 0
6                                6256                 1                 0
7                                6228                 1                 0
8                                6244                 1                 0
9                                6230                 1                 0

   Wilderness_Area3  Wilderness_Area4  Soil_Type1  Soil_Type2  Soil_Type3  \
0                 0                 0           0           0           0
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
7                 0                 0           0           0           0
8                 0                 0           0           0           0
9                 0                 0           0           0           0

   Soil_Type4  Soil_Type5  Soil_Type6  Soil_Type7  Soil_Type8  Soil_Type9  \
0           0           0           0           0           0           0
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
7           0           0           0           0           0           0
8           0           0           0           0           0           0
9           0           0           0           0           0           0

   Soil_Type10  Soil_Type11  Soil_Type12  Soil_Type13  Soil_Type14  \
0            0            0            0            0            0
1            0            0            0            0            0
```

```
2            0            0            1            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
7            0            0            0            0            0
8            0            0            0            0            0
9            0            0            0            0            0

   Soil_Type15  Soil_Type16  Soil_Type17  Soil_Type18  Soil_Type19  \
0            0            0            0            0            0
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
7            0            0            0            0            0
8            0            0            0            0            0
9            0            0            0            0            0

   Soil_Type20  Soil_Type21  Soil_Type22  Soil_Type23  Soil_Type24  \
0            0            0            0            0            0
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
7            0            0            0            0            0
8            0            0            0            0            0
9            0            0            0            0            0

   Soil_Type25  Soil_Type26  Soil_Type27  Soil_Type28  Soil_Type29  \
0            0            0            0            0            1
1            0            0            0            0            1
2            0            0            0            0            0
3            0            0            0            0            0
4            0            0            0            0            1
5            0            0            0            0            1
6            0            0            0            0            1
7            0            0            0            0            1
8            0            0            0            0            1
9            0            0            0            0            1

   Soil_Type30  Soil_Type31  Soil_Type32  Soil_Type33  Soil_Type34  \
0            0            0            0            0            0
1            0            0            0            0            0
```

```
2             0             0             0             0             0
3             1             0             0             0             0
4             0             0             0             0             0
5             0             0             0             0             0
6             0             0             0             0             0
7             0             0             0             0             0
8             0             0             0             0             0
9             0             0             0             0             0

   Soil_Type35  Soil_Type36  Soil_Type37  Soil_Type38  Soil_Type39  \
0             0             0             0             0             0
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
7             0             0             0             0             0
8             0             0             0             0             0
9             0             0             0             0             0

   Soil_Type40  Cover_Type
0             0           5
1             0           5
2             0           2
3             0           2
4             0           5
5             0           2
6             0           5
7             0           5
8             0           5
9             0           5
```

In [7]:
```
#pd.set_option('display.max_columns', None)
train.describe()
```

Out[7]:
```
                Id     Elevation        Aspect         Slope  \
count  15120.00000  15120.000000  15120.000000  15120.000000
mean    7560.50000   2749.322553    156.676653     16.501587
std     4364.91237    417.678187    110.085801      8.453927
min        1.00000   1863.000000      0.000000      0.000000
25%     3780.75000   2376.000000     65.000000     10.000000
50%     7560.50000   2752.000000    126.000000     15.000000
75%    11340.25000   3104.000000    261.000000     22.000000
max    15120.00000   3849.000000    360.000000     52.000000

       Horizontal_Distance_To_Hydrology  Vertical_Distance_To_Hydrology  \
count                      15120.000000                    15120.000000
```

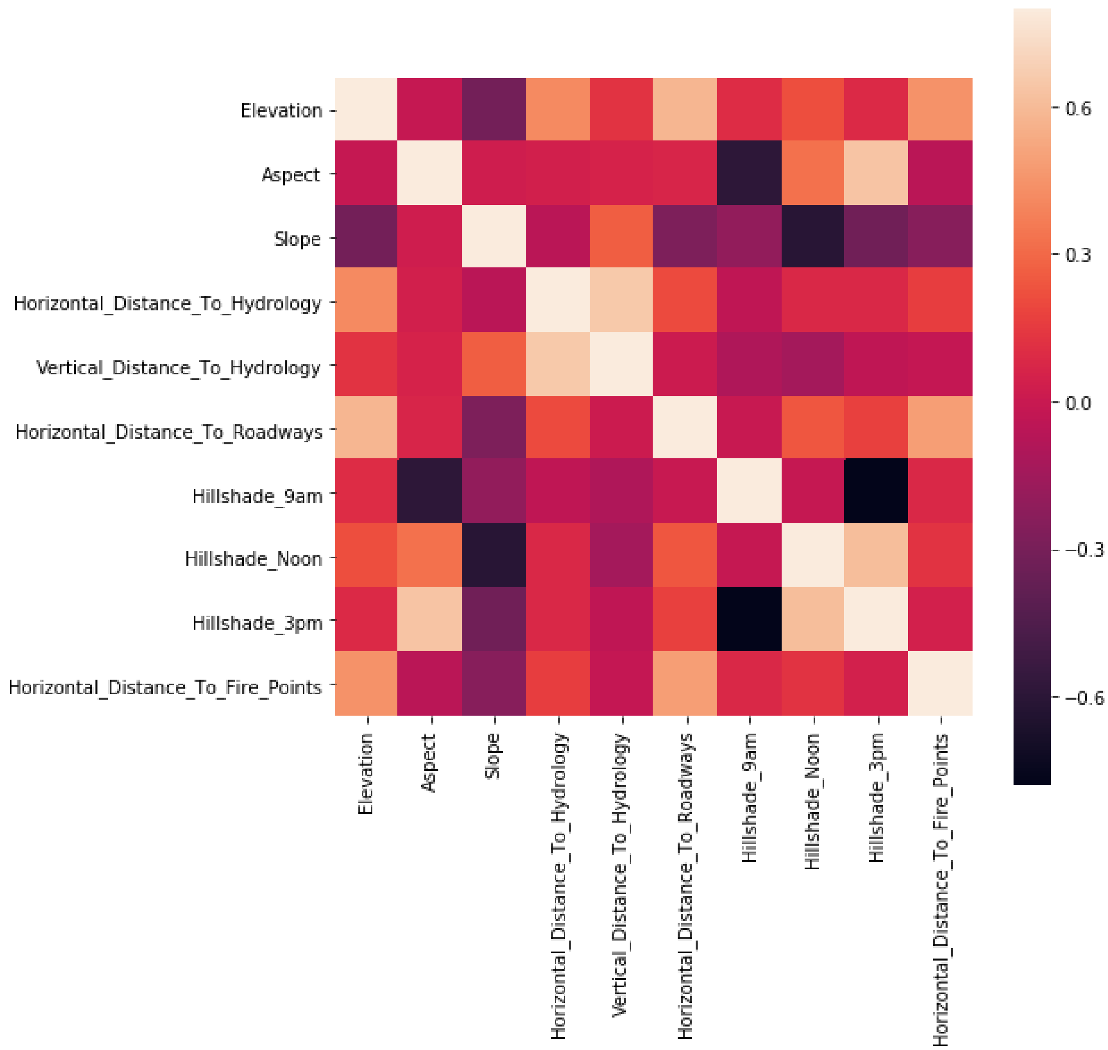
```
mean                 227.195701                           51.076521
std                  210.075296                           61.239406
min                    0.000000                         -146.000000
25%                   67.000000                            5.000000
50%                  180.000000                           32.000000
75%                  330.000000                           79.000000
max                 1343.000000                          554.000000

       Horizontal_Distance_To_Roadways  Hillshade_9am  Hillshade_Noon  \
count                     15120.000000   15120.000000    15120.000000
mean                       1714.023214     212.704299      218.965608
std                        1325.066358      30.561287       22.801966
min                           0.000000       0.000000       99.000000
25%                         764.000000     196.000000      207.000000
50%                        1316.000000     220.000000      223.000000
75%                        2270.000000     235.000000      235.000000
max                        6890.000000     254.000000      254.000000

       Hillshade_3pm  Horizontal_Distance_To_Fire_Points  Wilderness_Area1  \
count   15120.000000                        15120.000000      15120.000000
mean      135.091997                         1511.147288          0.237897
std        45.895189                         1099.936493          0.425810
min         0.000000                            0.000000          0.000000
25%       106.000000                          730.000000          0.000000
50%       138.000000                         1256.000000          0.000000
75%       167.000000                         1988.250000          0.000000
max       248.000000                         6993.000000          1.000000

       Wilderness_Area2  Wilderness_Area3  Wilderness_Area4    Soil_Type1  \
count      15120.000000      15120.000000      15120.000000  15120.000000
mean           0.033003          0.419907          0.309193      0.023479
std            0.178649          0.493560          0.462176      0.151424
min            0.000000          0.000000          0.000000      0.000000
25%            0.000000          0.000000          0.000000      0.000000
50%            0.000000          0.000000          0.000000      0.000000
75%            0.000000          1.000000          1.000000      0.000000
max            1.000000          1.000000          1.000000      1.000000

         Soil_Type2    Soil_Type3    Soil_Type4    Soil_Type5    Soil_Type6  \
count  15120.000000  15120.000000  15120.000000  15120.000000  15120.000000
mean       0.041204      0.063624      0.055754      0.010913      0.042989
std        0.198768      0.244091      0.229454      0.103896      0.202840
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000      0.000000      0.000000
75%        0.000000      0.000000      0.000000      0.000000      0.000000
max        1.000000      1.000000      1.000000      1.000000      1.000000
```

```
       Soil_Type7   Soil_Type8   Soil_Type9  Soil_Type10  Soil_Type11  \
count     15120.0 15120.000000 15120.000000 15120.000000 15120.000000
mean          0.0     0.000066     0.000661     0.141667     0.026852
std           0.0     0.008133     0.025710     0.348719     0.161656
min           0.0     0.000000     0.000000     0.000000     0.000000
25%           0.0     0.000000     0.000000     0.000000     0.000000
50%           0.0     0.000000     0.000000     0.000000     0.000000
75%           0.0     0.000000     0.000000     0.000000     0.000000
max           0.0     1.000000     1.000000     1.000000     1.000000

        Soil_Type12   Soil_Type13   Soil_Type14 Soil_Type15   Soil_Type16  \
count  15120.000000  15120.000000  15120.000000     15120.0  15120.000000
mean       0.015013      0.031481      0.011177         0.0      0.007540
std        0.121609      0.174621      0.105133         0.0      0.086506
min        0.000000      0.000000      0.000000         0.0      0.000000
25%        0.000000      0.000000      0.000000         0.0      0.000000
50%        0.000000      0.000000      0.000000         0.0      0.000000
75%        0.000000      0.000000      0.000000         0.0      0.000000
max        1.000000      1.000000      1.000000         0.0      1.000000

        Soil_Type17   Soil_Type18   Soil_Type19   Soil_Type20   Soil_Type21  \
count  15120.000000  15120.000000  15120.000000  15120.000000  15120.000000
mean       0.040476      0.003968      0.003042      0.009193      0.001058
std        0.197080      0.062871      0.055075      0.095442      0.032514
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000      0.000000      0.000000
75%        0.000000      0.000000      0.000000      0.000000      0.000000
max        1.000000      1.000000      1.000000      1.000000      1.000000

        Soil_Type22   Soil_Type23   Soil_Type24   Soil_Type25   Soil_Type26  \
count  15120.000000  15120.000000  15120.000000  15120.000000  15120.000000
mean       0.022817      0.050066      0.016997      0.000066      0.003571
std        0.149326      0.218089      0.129265      0.008133      0.059657
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000      0.000000      0.000000
75%        0.000000      0.000000      0.000000      0.000000      0.000000
max        1.000000      1.000000      1.000000      1.000000      1.000000

        Soil_Type27   Soil_Type28   Soil_Type29   Soil_Type30   Soil_Type31  \
count  15120.000000  15120.000000  15120.000000  15120.000000  15120.000000
mean       0.000992      0.000595      0.085384      0.047950      0.021958
std        0.031482      0.024391      0.279461      0.213667      0.146550
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000      0.000000      0.000000
75%        0.000000      0.000000      0.000000      0.000000      0.000000
```

```
max        1.000000     1.000000     1.000000     1.000000     1.000000

         Soil_Type32  Soil_Type33  Soil_Type34  Soil_Type35  Soil_Type36  \
count   15120.000000 15120.000000 15120.000000 15120.000000 15120.000000
mean        0.045635     0.040741     0.001455     0.006746     0.000661
std         0.208699     0.197696     0.038118     0.081859     0.025710
min         0.000000     0.000000     0.000000     0.000000     0.000000
25%         0.000000     0.000000     0.000000     0.000000     0.000000
50%         0.000000     0.000000     0.000000     0.000000     0.000000
75%         0.000000     0.000000     0.000000     0.000000     0.000000
max         1.000000     1.000000     1.000000     1.000000     1.000000

         Soil_Type37  Soil_Type38  Soil_Type39  Soil_Type40   Cover_Type
count   15120.000000 15120.000000 15120.000000 15120.000000 15120.000000
mean        0.002249     0.048148     0.043452     0.030357     4.000000
std         0.047368     0.214086     0.203880     0.171574     2.000066
min         0.000000     0.000000     0.000000     0.000000     1.000000
25%         0.000000     0.000000     0.000000     0.000000     2.000000
50%         0.000000     0.000000     0.000000     0.000000     4.000000
75%         0.000000     0.000000     0.000000     0.000000     6.000000
max         1.000000     1.000000     1.000000     1.000000     7.000000
```

In [8]:
```
#train = train.drop(['Soil_Type7', 'Soil_Type15'], axis = 1)
#test = test.drop(['Soil_Type7', 'Soil_Type15'], axis = 1)

#Drop 'id'  iloc[row,col]
train=train.iloc[:,1:]
test=test.iloc[:,1:]
```

In [11]:
```
corrmat = train.iloc[:,:10].corr()
ax = plt.subplots(figsize = (8,8))
sns.heatmap(corrmat,vmax=0.8,square=True);
```



In [14]:
```
size=10
data=train.iloc[:,:size]
cols = data.columns
data_corr=data.corr()
threshold=0.5
corr_list=[]
for i in range(0, 10):
    for j in range(i+1, 10):
        if data_corr.iloc[i,j]>= threshold and data_corr.iloc[i,j]<1\
        or data_corr.iloc[i,j] <0 and data_corr.iloc[i,j]<=-threshold:
            corr_list.append([data_corr.iloc[i,j],i,j])

corr_list
```

Out[14]:
```
[[0.5786589907340067, 0, 5],
 [-0.5939974281313112, 1, 6],
```

```
 [0.635022364019874, 1, 8],
 [-0.6126128724172692, 2, 7],
 [0.6521424712357364, 3, 4],
 [-0.779964742447544, 6, 8],
 [0.6145263872475779, 7, 8]]
```

In [10]:
```python
s_corr_list = sorted(corr_list,key= lambda x: -abs(x[0]))

# print the higher values
for v,i,j in s_corr_list:
    print("%s and %s = %.2f" % (cols[i], cols[j], v))
```

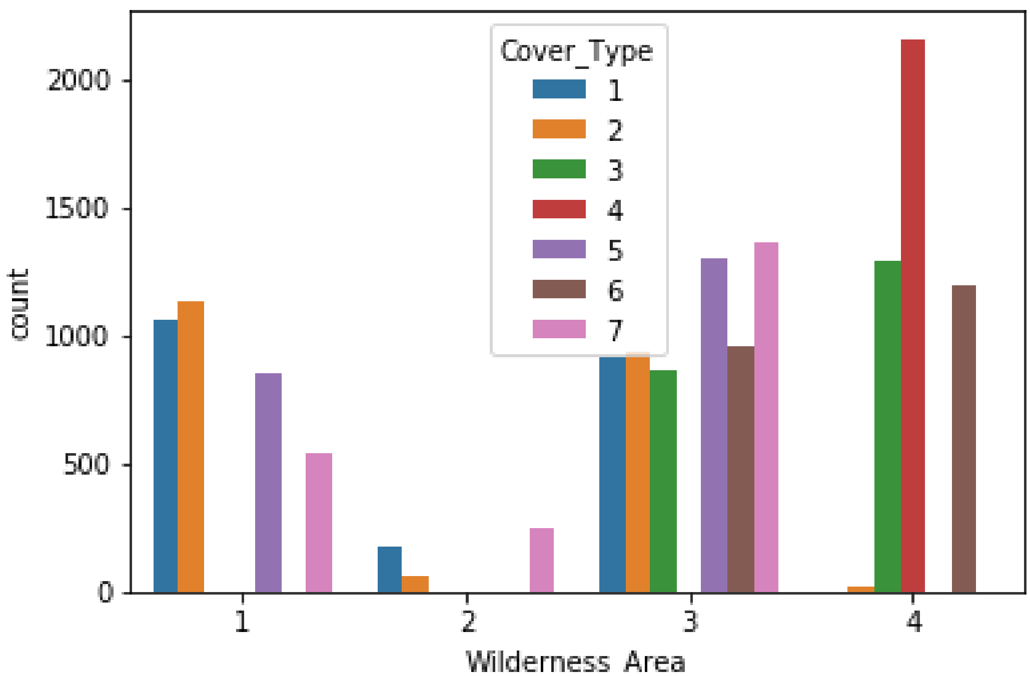
```
Hillshade_9am and Hillshade_3pm = -0.78
Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology = 0.65
Aspect and Hillshade_3pm = 0.64
Hillshade_Noon and Hillshade_3pm = 0.61
Slope and Hillshade_Noon = -0.61
Aspect and Hillshade_9am = -0.59
Elevation and Horizontal_Distance_To_Roadways = 0.58
```

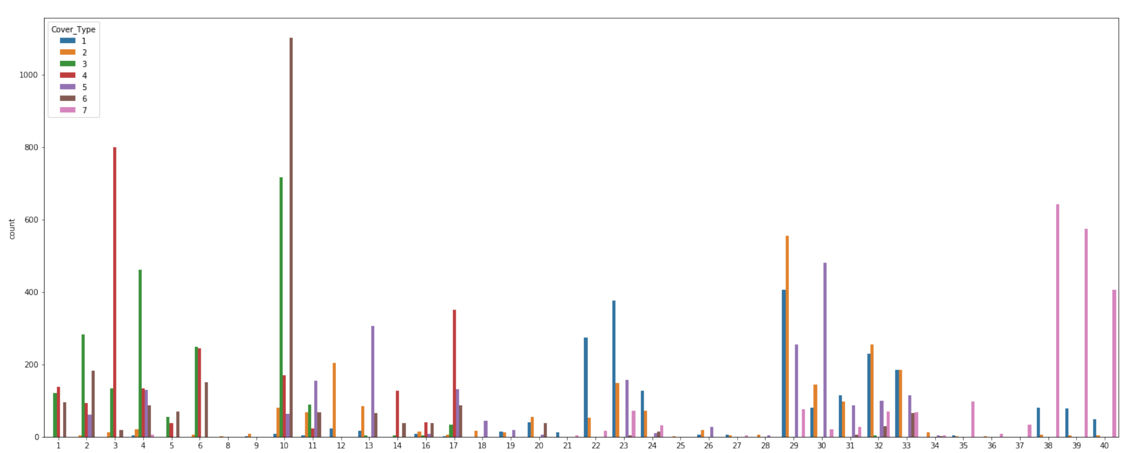
In [11]:
```python
train.Wilderness_Area2.value_counts()
```

Out[11]:
```
0    14621
1      499
Name: Wilderness_Area2, dtype: int64
```

In [15]:
```python
# Group one-hot encoded variables of a category into one single variable
cols = train.columns
r,c = train.shape

# Create a new dataframe with r rows, one column for each encoded category, and targe
new_data = pd.DataFrame(index= np.arange(0,r), columns=['Wilderness_Area', 'Soil_Type

# Make an entry in data for each r for category_id, target_value
for i in range(0,r):
    p = 0;
    q = 0;
    # Category1_range
    for j in range(10,14):
        if (train.iloc[i,j] == 1):
            p = j-9 # category_class
            break
    # Category2_range
    for k in range(14,54):
        if (train.iloc[i,k] == 1):
            q = k-13 # category_class
            break
    # Make an entry in data for each r
```

```
        new_data.iloc[i] = [p,q,train.iloc[i, c-1]]

# plot for category1
sns.countplot(x = 'Wilderness_Area', hue = 'Cover_Type', data = new_data)
plt.show()

# Plot for category2
plt.rc("figure", figsize = (25,10))
sns.countplot(x='Soil_Type', hue = 'Cover_Type', data= new_data)
plt.show()
```





In [13]:
```
#check normality of non-binary variables
train.iloc[:,:10].skew()
```

Out[13]:
```
Elevation                             0.075640
Aspect                                0.450935
Slope                                 0.523658
Horizontal_Distance_To_Hydrology      1.488052
Vertical_Distance_To_Hydrology        1.537776
Horizontal_Distance_To_Roadways       1.247811
Hillshade_9am                        -1.093681
Hillshade_Noon                       -0.953232
Hillshade_3pm                        -0.340827
Horizontal_Distance_To_Fire_Points    1.617099
dtype: float64
```

In [14]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

r,c = train.shape
X_train = train.iloc[:,:c-1]
y_train = train["Cover_Type"]


# Setting parameters
x_data, x_test_data, y_data, y_test_data = train_test_split(train, y_train, test_size
rf_para = [{'n_estimators':[50, 100], 'max_depth':[5,10,15], 'max_features':[0.1, 0.3
           'min_samples_leaf':[1,3], 'bootstrap':[True, False]}]
```

In [15]:
```
from sklearn.model_selection  import GridSearchCV, RandomizedSearchCV
rfc = GridSearchCV(RandomForestClassifier(), param_grid=rf_para, cv = 10, n_jobs=-1)
rfc.fit(x_data, y_data)
rfc.best_params_
```

Out[15]:
```
{'bootstrap': True,
 'max_depth': 15,
 'max_features': 0.3,
 'min_samples_leaf': 1,
 'n_estimators': 50}
```

In [16]:
```
RFC = RandomForestClassifier(n_estimators=100, max_depth=15, max_features=0.3, bootst
                             n_jobs=-1)
RFC.fit(X_train, y_train)
rfc_pred=RFC.predict(test)
```

In [39]:
```
solution = pd.DataFrame({'Id':Test.Id, 'Cover_Type':rfc_pred}, columns = ['Id','Cover
solution.to_csv('rfc_sol.csv', index=False)
```

In [ ]: