Lab 9

Jiarong Ye

October 19, 2018

Import packages

```
In [23]: import datascience
         from datascience import *
         import numpy as np
         import pandas as pd
         import graphviz
         from sklearn.model_selection import train_test_split
         from sklearn.pipeline import Pipeline
         from sklearn import tree
         from sklearn.tree import DecisionTreeClassifier
         from sklearn import metrics
```

Read in the dataset with pandas

```
In [63]: Ben_Pass= pd.read_csv('Ben-NE-9-10-2015-pass-6.csv', sep=",")
         Ben_Pass_2 = pd.read_csv('Ben-NE-9-10-2015-pass-4-1.csv', sep=',')
```

```
In [64]: Ben_Pass.head()
```

	down	ydstogo	Yards.Gained.PrevPlay	AirYards	PassLocation	PassOutcome
0	1	10	18	-4	1	1
1	1	10	0	9	1	1
2	3	22	6	1	1	1
3	1	10	0	7	-1	1
4	1	10	13	6	-1	1

```
In [25]: Ben_Pass_2.head()
```

	down	ydstogo	Yards.Gained.PrevPlay	AirYards	PassLocation	PassOutcome
0	1	10	18	-4	1	1
1	1	10	0	9	1	1
2	3	22	6	1	1	1
3	1	10	0	7	-1	1
4	1	10	13	6	-1	1

Conduct Correlation Analysis

```
In [65]: Ben_Pass.corr(method='pearson')
```

	down	ydstogo	Yards.PrevPlay	AirYards	PassLocation	PassOutcome
down	1.000000	-0.293906	-0.306207	-0.032894	-0.060043	-0.076448
ydstogo	-0.293906	1.000000	-0.060572	0.091813	-0.058176	0.249902
Yards.PrevPlay	-0.306207	-0.060572	1.000000	0.022748	-0.081227	0.144037
AirYards	-0.032894	0.091813	0.022748	1.000000	0.056054	-0.286445
PassLocation	-0.060043	-0.058176	-0.081227	0.056054	1.000000	-0.223061
PassOutcome	-0.076448	0.249902	0.144037	-0.286445	-0.223061	1.000000

```
In [27]: Ben_Pass_2.corr(method='pearson')
```

	down	ydstogo	Yards.PrevPlay	Yards.Gained	AirYards	PassLocation	PassOutcome
down	1.000000	-0.293906	-0.306207	-0.054182	-0.032894	-0.060043	-0.076448
ydstogo	-0.293906	1.000000	-0.060572	0.222043	0.091813	-0.058176	0.249902
Yards.PrevPlay	-0.306207	-0.060572	1.000000	0.131102	0.022748	-0.081227	0.144037
Yards.Gained	-0.054182	0.222043	0.131102	1.000000	0.363336	0.019662	0.537075
AirYards	-0.032894	0.091813	0.022748	0.363336	1.000000	0.056054	-0.286445
PassLocation	-0.060043	-0.058176	-0.081227	0.019662	0.056054	1.000000	-0.223061
PassOutcome	-0.076448	0.249902	0.144037	0.537075	-0.286445	-0.223061	1.000000

Build Decision Trees

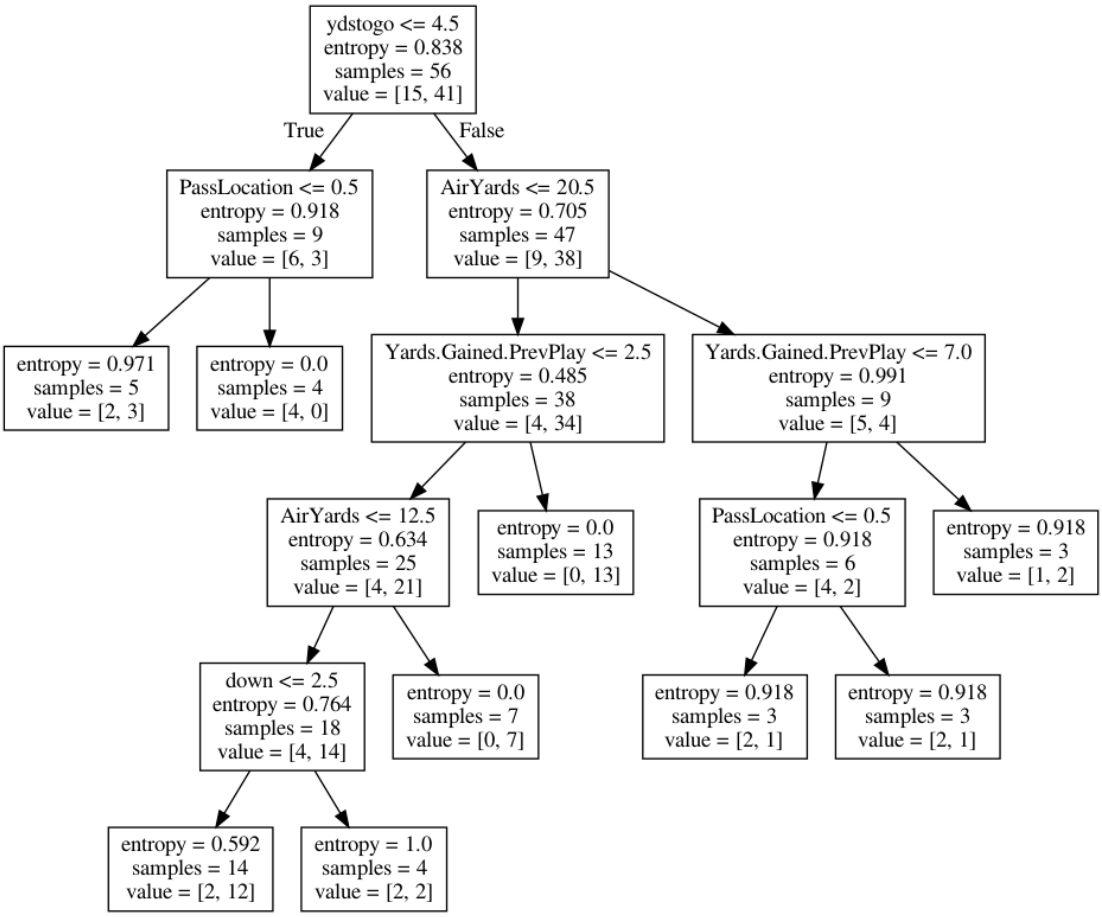
```
In [113]: def evaluate(dataset, seed, arg):
              X = dataset.values[:, :-1]
              y = dataset.values[:, -1]
              X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state =
              clf = DecisionTreeClassifier(criterion = 'entropy', random_state = 100,
                                           max_depth=5, min_samples_leaf=3)
              clf.fit(X_train, y_train)
              dot_data= tree.export_graphviz(clf, out_file=None, feature_names=dataset.columns[:-1])
              graph = graphviz.Source(dot_data)
              graph.render('BenPass{}'.format(str(arg)))
              predicted_completion = clf.predict(X_test)
              print('Ben Pass dataset {} classification report: \n'.format(str(arg)),
               metrics.classification_report(y_true=y_test, y_pred=predicted_completion))

In [114]: evaluate(Ben_Pass, 2018, 1)
          evaluate(Ben_Pass_2, 2018, 2)

Ben Pass dataset 1 classification report:
              precision    recall  f1-score   support

           0       0.75      0.75      0.75         4
           1       0.90      0.90      0.90        10
```

```
   micro avg       0.86      0.86      0.86        14
   macro avg       0.82      0.82      0.82        14
weighted avg       0.86      0.86      0.86        14

Ben Pass dataset 2 classification report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00        10

   micro avg       1.00      1.00      1.00        14
   macro avg       1.00      1.00      1.00        14
weighted avg       1.00      1.00      1.00        14
```

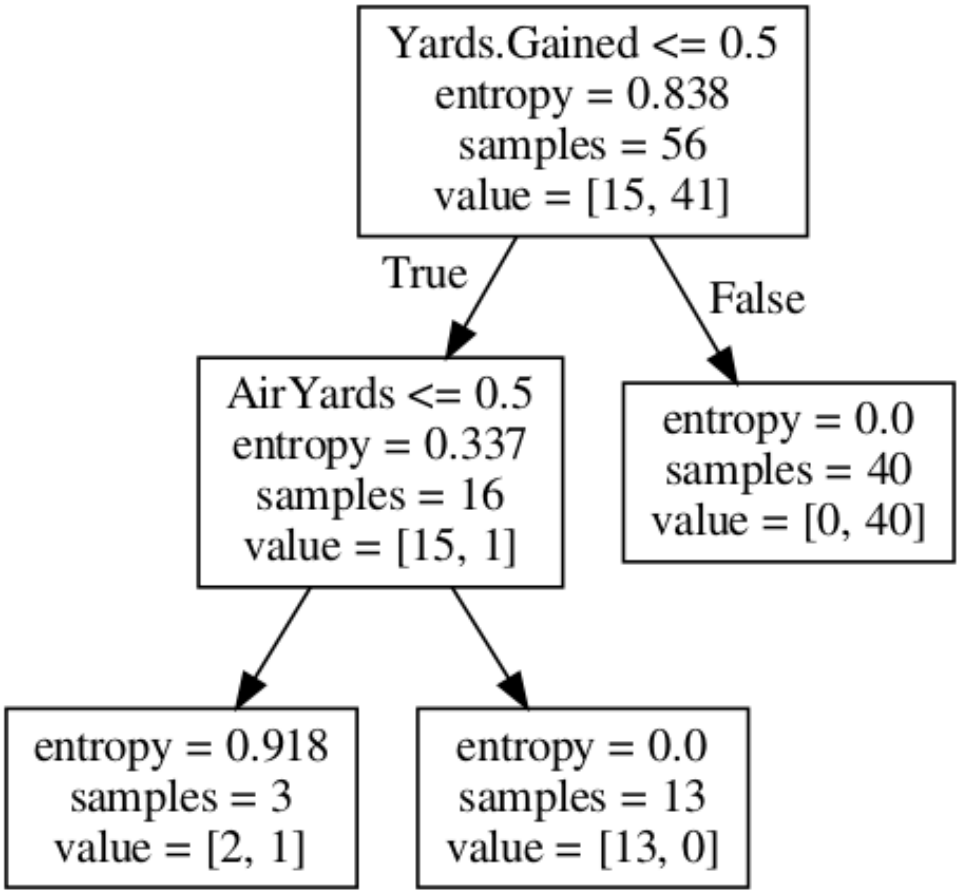
So from the classification report we could see that the Decision Tree built for the second dataset (**with Yard.Gained**) has no misclassified instances, reaching 100% accuracy score.

Discussion of the Correlation and DecisionTree

Ben_Pass_1 (Without the `Yard.Gained` variable)



Ben_Pass_2 (With the `Yard.Gained` variable)



The second Ben Pass dataset that includes the `Yards.Gained` variable stop after two splits (if the random state is changed to another number, then the tree actually stops after one single split), the reason for which is:

Revisit the correlation matrix:

```
In [87]: Ben_Pass_2.corr(method='pearson')
```

	down	ydstogo	Yards.PrevPlay	Yards.Gained	AirYards	PassLocation	PassOutcome
down	1.000000	-0.293906	-0.306207	-0.054182	-0.032894	-0.060043	-0.076448
ydstogo	-0.293906	1.000000	-0.060572	0.222043	0.091813	-0.058176	0.249902
Yards.PrevPlay	-0.306207	-0.060572	1.000000	0.131102	0.022748	-0.081227	0.144037
Yards.Gained	-0.054182	0.222043	0.131102	1.000000	0.363336	0.019662	0.537075
AirYards	-0.032894	0.091813	0.022748	0.363336	1.000000	0.056054	-0.286445
PassLocation	-0.060043	-0.058176	-0.081227	0.019662	0.056054	1.000000	-0.223061
PassOutcome	-0.076448	0.249902	0.144037	0.537075	-0.286445	-0.223061	1.000000

Conclusion:

from the correlation matrix above, we could observe that the correlation coefficient between *Yards.Gained* and *PassOutcome* is 0.537075 (positively correlated), thus we should be able to conclude that *Yards.Gained* has a significant influence on the *PassOutcome*. The more correlated a variable is to the target variable, the higher up it should be in the constructed Decision Tree, hence explains that the *Yards.Gained* is the root node, and the second largest correlation coefficient among the other variables to *PassOutcome* is *AirYards* as 0.286445 (negatively correlated), so the first child right after the root node is *AirYards*.