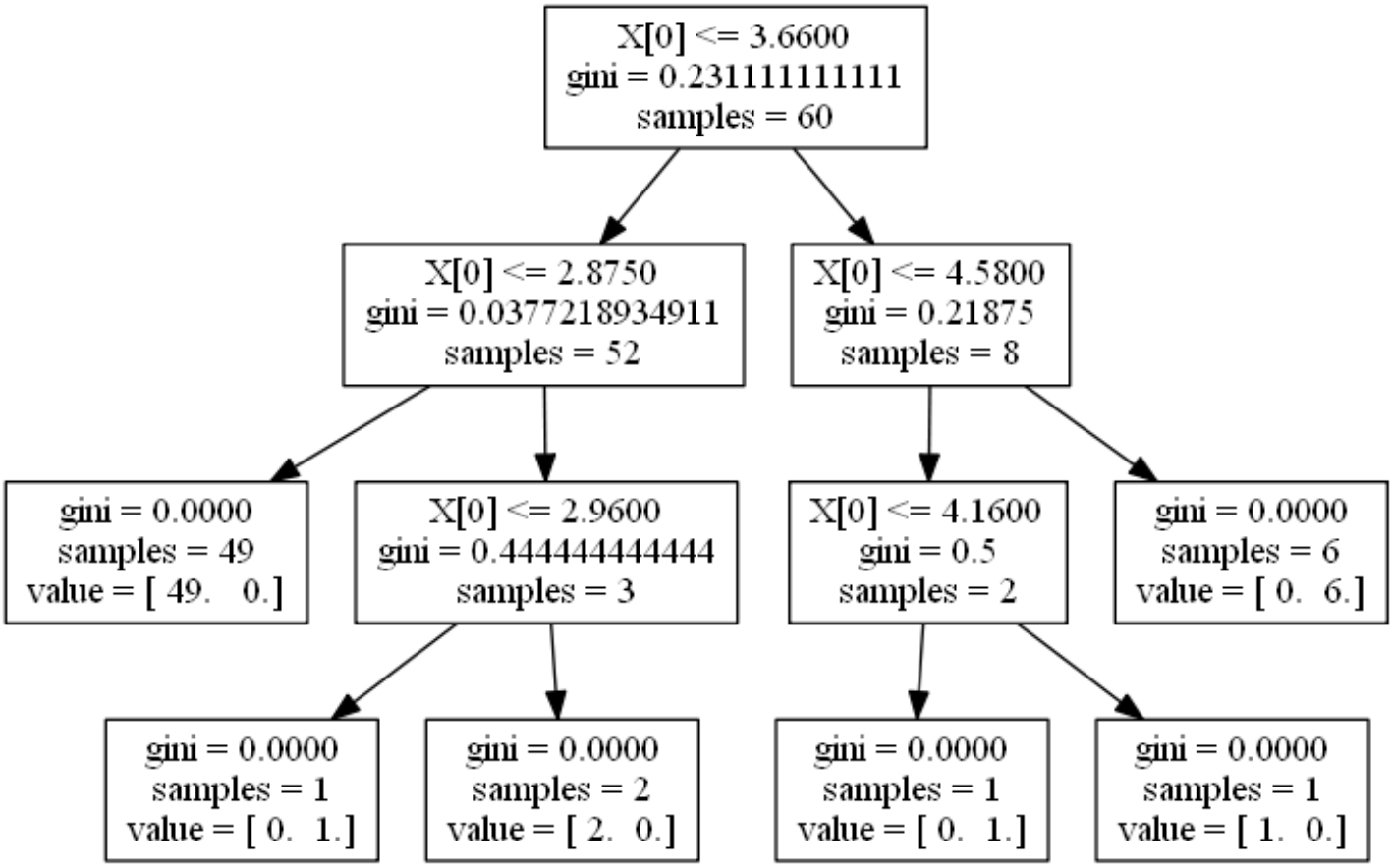
Real life data science

JOSE RAMON PASILLAS DIAZ




```
from pandas import Series,DataFrame
import pandas as pd
import numpy as np
from sklearn.cross_validation import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report
import sklearn.metrics

mars = pd.read_csv('marscrater_pds.csv', low_memory=False)
# Taka a random sample od the data
rows = np.random.choice(mars.index.values, 100)
```

```
sampled_mars = mars.ix[rows]
mars_clean=sampled_mars.dropna()
mars_clean.dtypes
mars_clean.describe()
# binnig response variable into two categories
mars_clean[“DEPTH”]=pd.cut(sampled_mars.DEPTH_RIMFLOOR_TOPOG,
[-5,sampled_mars[“DEPTH_RIMFLOOR_TOPOG”].mean(),1171],labels=[0,1])
mars_clean[“DEPTH”] = mars_clean[“DEPTH”].astype('category’)
# set explanatory and response variables
predictors=mars_clean[['DIAM_CIRCLE_IMAGE’,'NUMBER_LAYERS’]]
target=mars_clean.DEPTH
# Set train and test set
pred_train,pred_test,tar_train,tar_test=train_test_split(predictors,target,test_size=.4)
#Building model on training data
classifier=DecisionTreeClassifier()
classifier=classifier.fit(pred_train,tar_train)
predictions=classifier.predict(pred_test)
sklearn.metrics.confusion_matrix(tar_test,predictions)
sklearn.metrics.accuracy_score(tar_test,predictions)
#Displaying the decision tree
from sklearn import tree
#from StringIO import StringIO
from io import StringIO
#from StringIO import StringIO
from IPython.display import Image
out = StringIO()
tree.export_graphviz(classifier, out_file=out)
import pydotplus
graph=pydotplus.graph_from_dot_data(out.getvalue())
Image(graph.create_png())
```

In this post we will use classification trees to test for nonlinear relationship between our response variable craters depth and the explanatory variables diameter and number of layers. Due to our limitations in hardware we sampled the original dataset to have a random sample that produces an easy to process and interpret tree. Note that we could have used a larger sample, but for the purposes of this assignment we believe that interpretability is a key factor.

Our final decision tree has correctly classified about 95% of the craters diameters only using the first of the explanatory variables: diameter.

Feb 2nd, 2016

MORE YOU MIGHT LIKE

Show more