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

In [15]:
```
df = pd.read_csv('E:/adult-income-usa/adult-income-usa-cleaned.csv')
df.head()
```

Out[15]:

	row	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week	workclass	marital-status	race	Ma
0	0	25	226802	7	0	0	40	0	0	0	1
1	1	38	89814	9	0	0	50	0	1	1	1
2	2	28	336951	12	0	0	40	1	1	1	1
3	3	44	160323	10	7688	0	40	0	1	0	1
4	5	34	198693	6	0	0	30	0	0	1	1

In [16]:
```
x = df.drop(['>50K', 'row'], axis=1)
y = df['>50K']
```

In [17]:
```
from sklearn.cross_validation import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3)
```

In [7]:
```
from sklearn.linear_model import LogisticRegression
```

In [18]:
```
logmodel = LogisticRegression()
```

In [19]:
```
logmodel.fit(x_train, y_train)
```

Out[19]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

In [20]:
```
predictions = logmodel.predict(x_test)
```

In [11]:
```
from sklearn.metrics import classification_report
```

In [21]:
```
print (classification_report(y_test, predictions))
```

```
             precision    recall  f1-score   support

          0       0.80      0.96      0.88     10239
          1       0.71      0.27      0.39      3328

avg / total       0.78      0.79      0.76     13567
```

In [13]:
```
from sklearn.metrics import confusion_matrix
```

In [22]:
```
confusion_matrix(y_test, predictions)
```

Out[22]:
```
array([[9876,  363],
       [2438,  890]], dtype=int64)
```