Test run of data.csv

autoML.py defaults:

```
usage: autoML.py [-h] [-m MODEL_TYPE] [-i INPUT_FILE] [-file2 SECONDARY_FILE]
                 [-w SECONDARY_WEIGHTS] [-d DISTANCEFN] [-sprs SPARSITY]
                 [-r RADIUS] [-t MAX_TIME] [-n MAX_ITERATIONS] [-pca N_PCA]
                 [-e N_EXPERTS] [-s {none,manual,auto,manual_both,auto_both}]
                 [-f HIERARCHY_FOLDER]

optional arguments:
  -h, --help            show this help message and exit
  -m MODEL_TYPE, --model_type MODEL_TYPE
                        choose classification(default), regression, clustering
                        or outlier_detection
  -i INPUT_FILE, --input_file INPUT_FILE
                        primary input file to be analyzed (default=data.csv)
  -file2 SECONDARY_FILE, --secondary_file SECONDARY_FILE
                        optional secondary input file, triggers multi-dataset
                        analysis (default=None)
  -w SECONDARY_WEIGHTS, --secondary_weights SECONDARY_WEIGHTS
                        weights for features in secondary file, default=False
  -d DISTANCEFN, --distanceFn DISTANCEFN
                        choose L_1Norm(n), euclidean(n), L_infinityNorm(n),
                        distanceOnEarth(n), L_1Norm_cat(n), or
                        L_infinityNorm_cat(n), where n=1,2,3,... is the chosen
                        dimension for calculating distances. Default is
                        L_1Norm(1)
  -sprs SPARSITY, --sparsity SPARSITY
                        sparsity threshold for including records in secondary
                        input file
  -r RADIUS, --radius RADIUS
                        radius for cutoff of the distance function (default=1)
  -t MAX_TIME, --max_time MAX_TIME
                        maximum time in seconds for training all models. The
                        default value is 1440 seconds.
  -n MAX_ITERATIONS, --max_iterations MAX_ITERATIONS
                        max iterations for cross-validation of each individual
                        model fit. The default is 10 for clustering, 100 for
                        classification and 100 for regression.
  -pca N_PCA, --n_pca N_PCA
                        number of PCA components for outlier detection.
                        (default is 4)
  -e N_EXPERTS, --n_experts N_EXPERTS
                        number of experts for Ensemble scoring. (default is 5)
  -s {none,manual,auto,manual_both,auto_both}, --privatize_data {none,manual,aut
o,manual_both,auto_both}
                        choose none, manual, or auto for privatization of the
                        data using ARX. For manual, an ARX window will launch.
                        For privatization of primary and secondary datasets,
                        choose manual_both or auto_both. Default is manual.
  -f HIERARCHY_FOLDER, --hierarchy_folder HIERARCHY_FOLDER
                        folder containing hierarchy files for sensitive data,
                        if provided by the user. Default is hierarchy
```

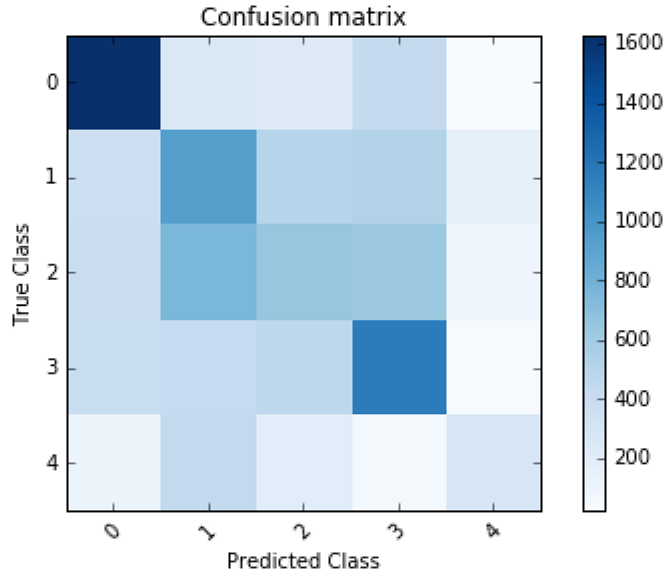
It currently handles four types of models: classification, regression, clustering and outlier detection. If the model type is classification or regression then the last column of the input data is assumed to be the dependent variable. Option to add a second dataset and a distance function: The distance function is used to assign elements of the second dataset to each row in the first dataset. A cutoff radius is used for the selection, with default initial value of 1. The -r option can be used to scale the distance function differently.

Output:

```
Python 3.5.2 |Anaconda 4.2.0 (64-bit)| (default, Jul 5 2016, 11:41:13) [MSC v.1900 64
bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 5.1.0 -- An enhanced Interactive Python.

In [4]: runfile('C:/Users/torres/Documents/GitHub/autoML-multiData/autoML.py',
wdir='C:/Users/torres/Documents/GitHub/autoML-multiData')
Reloaded modules: DeIdentify, DistanceFn, Model, FuseData, RandomizedSearchCluster,
TimeSeries, Experts, Data, Image, ClusterWrapper
Converting file to features
Dataset 'data.csv': (44106, 56)
Column names: ['acctype=1', 'acctype=10', 'acctype=11', 'acctype=12', 'acctype=13',
'acctype=14', 'acctype=15', 'acctype=2', 'acctype=3', 'acctype=4', 'acctype=5',
'acctype=6', 'acctype=7', 'acctype=8', 'acctype=9', 'acctype=99', 'lat', 'light=1',
'light=2', 'light=3', 'light=4', 'light=5', 'light=9', 'longitude', 'numvehs=1',
'numvehs=13-Aug', 'numvehs=2', 'numvehs=3', 'numvehs=5-Mar', 'numvehs=8-May',
'rdsurf=1', 'rdsurf=2', 'rdsurf=3', 'rdsurf=4', 'rdsurf=5', 'rdsurf=6', 'rdsurf=9',
'rodwycls=1', 'rodwycls=10', 'rodwycls=2', 'rodwycls=3', 'rodwycls=4', 'rodwycls=5',
'rodwycls=6', 'rodwycls=7', 'rodwycls=8', 'rodwycls=9', 'rodwycls=99', 'weather=1',
'weather=2', 'weather=3', 'weather=4', 'weather=5', 'weather=6', 'weather=7',
'weather=9']
Target name: severity
Target type: cat
Target classes: ['0' '1' '2' '3' '4']
Target encoding: [0 1 2 3 4]
Row 1: [ 0. 0. 1. ..., 0. 1. 0.] -> 0
Row -1: [ 0. 0. 0. ..., 0. 0. 0.] -> 4
GradientBoost
Time to fit 3 instances of GradientBoost: 11687.49s
```

```
AdaBoost
Time to fit 3 instances of AdaBoost: 57.20s
ExtraTrees
Time to fit 3 instances of ExtraTrees: 2.89s
LogisticRegression
Time to fit 3 instances of LogisticRegression: 122.51s
GaussianNB
Time to fit 3 instances of GaussianNB: 2.09s
DecisionTree
Time to fit 3 instances of DecisionTree: 6.18s
RandomForest
Time to fit 3 instances of RandomForest: 19.50s
SGD
Time to fit 3 instances of SGD: 3.78s
LDA
Time to fit 3 instances of LDA: 2.13s
KNeighbors
Time to fit 3 instances of KNeighbors: 19.39s
QDA
Time to fit 3 instances of QDA: 2.05s
Fitting GradientBoost (n_iterations=0, max_model_time=130s)
Skipping GradientBoost due to error: list index out of range
Fitting AdaBoost (n_iterations=6, max_model_time=130s)
Number of iterations: 6, Elapsed time: 186.30s
Fitting ExtraTrees (n_iterations=100, max_model_time=130s)
Number of iterations: 100, Elapsed time: 42.28s
Fitting LogisticRegression (n_iterations=3, max_model_time=130s)
Number of iterations: 3, Elapsed time: 20.80s
Fitting GaussianNB (n_iterations=100, max_model_time=130s)
Number of iterations: 100, Elapsed time: 51.86s
Fitting DecisionTree (n_iterations=63, max_model_time=130s)
Number of iterations: 63, Elapsed time: 77.04s
Fitting LDA (n_iterations=100, max_model_time=130s)
Number of iterations: 100, Elapsed time: 68.26s
Fitting SGD (n_iterations=100, max_model_time=130s)
Number of iterations: 100, Elapsed time: 86.68s
Fitting RandomForest (n_iterations=20, max_model_time=130s)
Number of iterations: 20, Elapsed time: 157.61s
Fitting KNeighbors (n_iterations=20, max_model_time=130s)
Number of iterations: 20, Elapsed time: 205.05s
Fitting QDA (n_iterations=100, max_model_time=130s)
Number of iterations: 100, Elapsed time: 57.81s

Number of models: 10
Models: ['RandomForest: 0.420423', 'ExtraTrees: 0.413168', 'SGD: 0.410356',
'AdaBoost: 0.409903', 'LogisticRegression: 0.409268', 'LDA: 0.407182', 'QDA:
0.388138', 'KNeighbors: 0.382969', 'DecisionTree: 0.378253', 'GaussianNB: 0.356398']


Ensemble Confusion Matrix (based on majority votes of top 5 models):
[[1629 248 222 426 20]
[ 368 926 496 521 152]
[ 385 757 653 625 87]
[ 393 419 463 1154 28]
[ 104 435 191 51 274]]
```

```
precision recall f1-score support

0 0.57 0.64 0.60 2545
1 0.33 0.38 0.35 2463
2 0.32 0.26 0.29 2507
3 0.42 0.47 0.44 2457
4 0.49 0.26 0.34 1055

avg / total 0.42 0.42 0.41 11027
```


Confusion matrix
True Class
Predicted Class