TPOT Cheat Sheet

TPOT (https://github.com/EpistasisLab/tpot) is a tool on-top of scikit-learn to automate optimization of machine learning pipelines using genetic programming.

Preprocessing required

Clean up the data (e.g. deal with missing values)

```
#use pandas or something like that for cleaning data
import pandas as pd
# do stuff!
```

Train-test-split

```
from sklearn.model_selection import train_test_split
train_test_split
```

TPOT classes

Import TPOT classifier or regressor

```
from tpot import TPOTClassifier
from tpot import TPOTRegressor

tpot = TPOTClassifier(*OPTIONS)
tpot = TPOTRegressor(*OPTIONS)
```

TPOT options

The options/arguments for both classes are identical.

Running TPOT using a timelimit

```
# time limit in minutes
tpot = TPOTClassifier(max_time_mins=60)
```

Options with explanation and default values

```
# continue previous pipeline or cold start from scratch
warm_start=False
# define number of cross-validation folds
cv=5
# optimization time per pipeline
max_eval_time_mins=5
# subsampling ratio (1.0: no subsampling)
subsample=1.0
# number of generations
generations=100
# number of individuals in population
population_size=100
# number of offsprings per generation population
offsprings_size=None #set number as int
# mutation rate (random changes per generation)
mutation_rate=0.9
# Python pseudo random number generator random_state
random_state=None
# number of CPU cores (incl. HT) used
n_jobs=1
# using dask
use_dask=False
# control output messages
verbosity=0 #1: minimal, 2: high, 3: all
```

Evaluation and save models

Evaluate model

```
# test set evaluation
tpo.score(X_test, y_test)
```

Export pipeline as a Python script

```
tpot.export(``filename.py'')
```