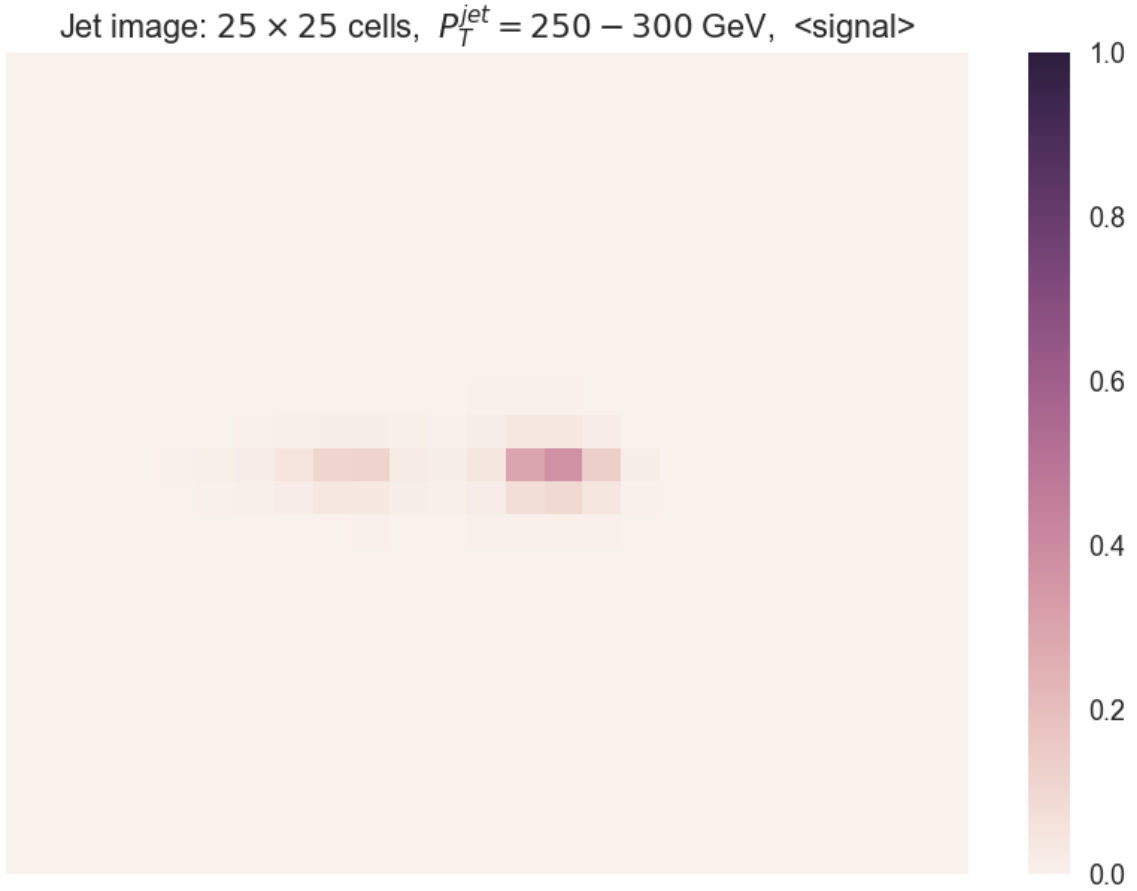
Boosted jet identification with Machine Learning

JOSÉ RUIZ

Recapitulation

Data generation and preprocessing

- Simulation
  - Pythia 8 + FastJet
- Preprocessing
  - ROOT
- Jet images
  - 25x25 calorimetric cells
  - Normalized energy between 0 - 1



Basic definitions

Signal

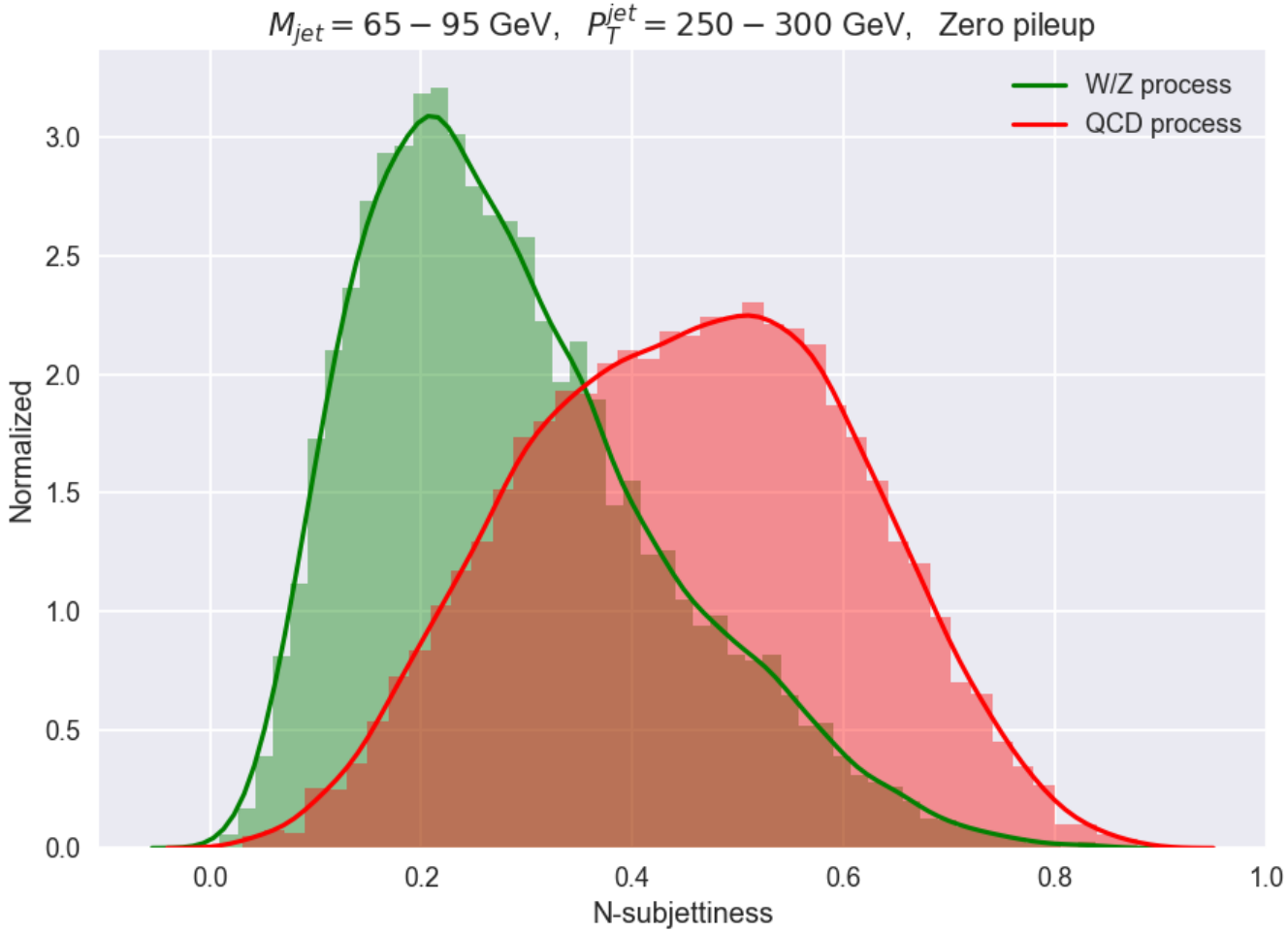
- High energy jets
- W/Z processes

Background

- QCD processes

N-subjettiness

- Substructure variable
- Discriminate signal and background



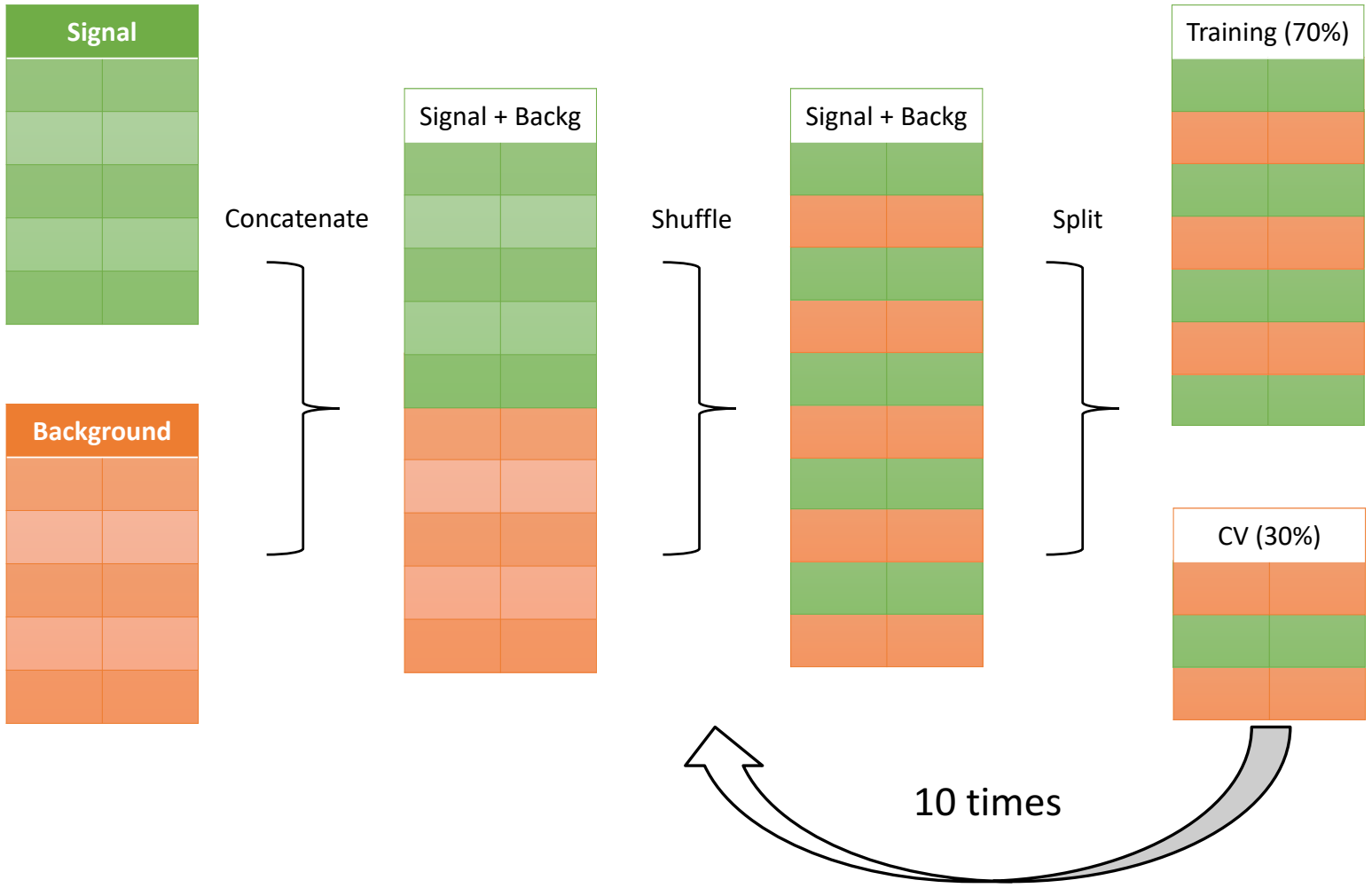
Model selection and evaluation

ML models

- Random Forest
- Logistic Regression
- Artificial Neural Networks
  - Multilayer Perceptron
  - Convolutional NN

Model evaluation

- Training a model and testing on the same data is a mistake
- Cross-validation (CV)
  - Random split into training and test sets



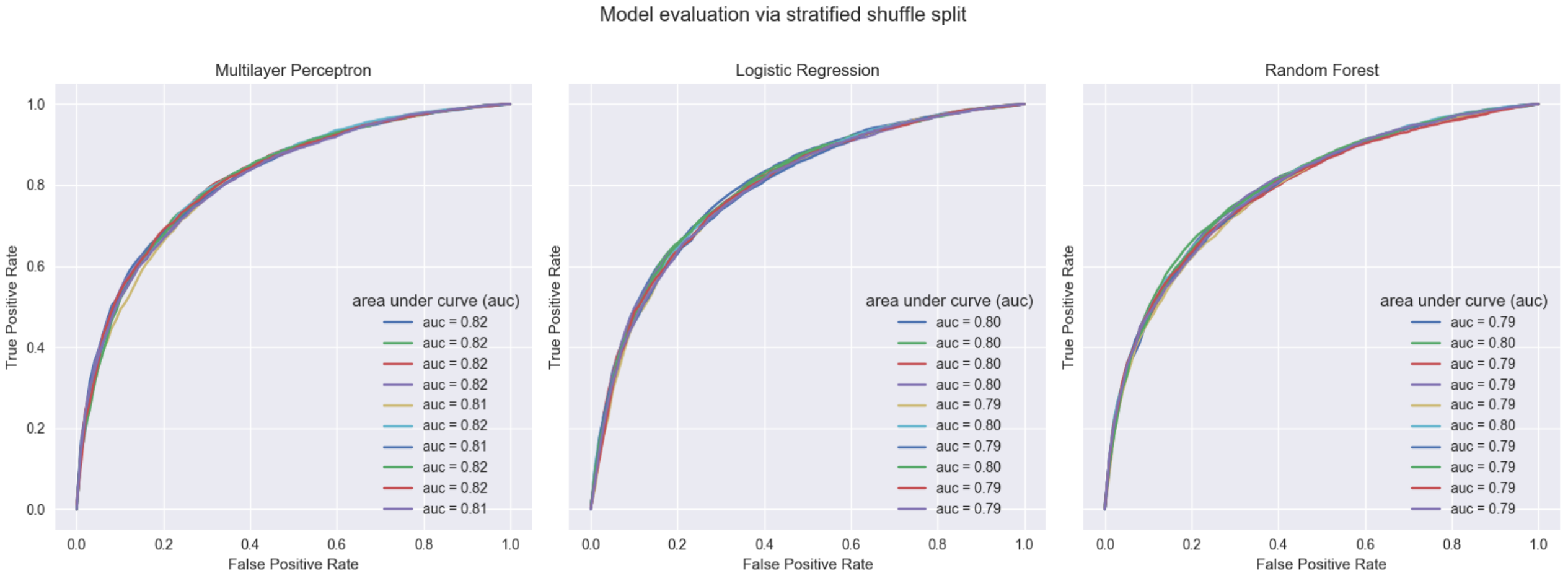
Stratified shuffle split

Stratified sampling

- Avoids the problem of imbalance in the distribution of the target classes

```
# Stratified shuffle split for cross validation
# Number of splits = 10
# Test sample = 30%
from sklearn.model_selection import StratifiedShuffleSplit
split = StratifiedShuffleSplit(n_splits=10, test_size=0.3, random_state=42)
```

Models performance



Convolutional Neural Net

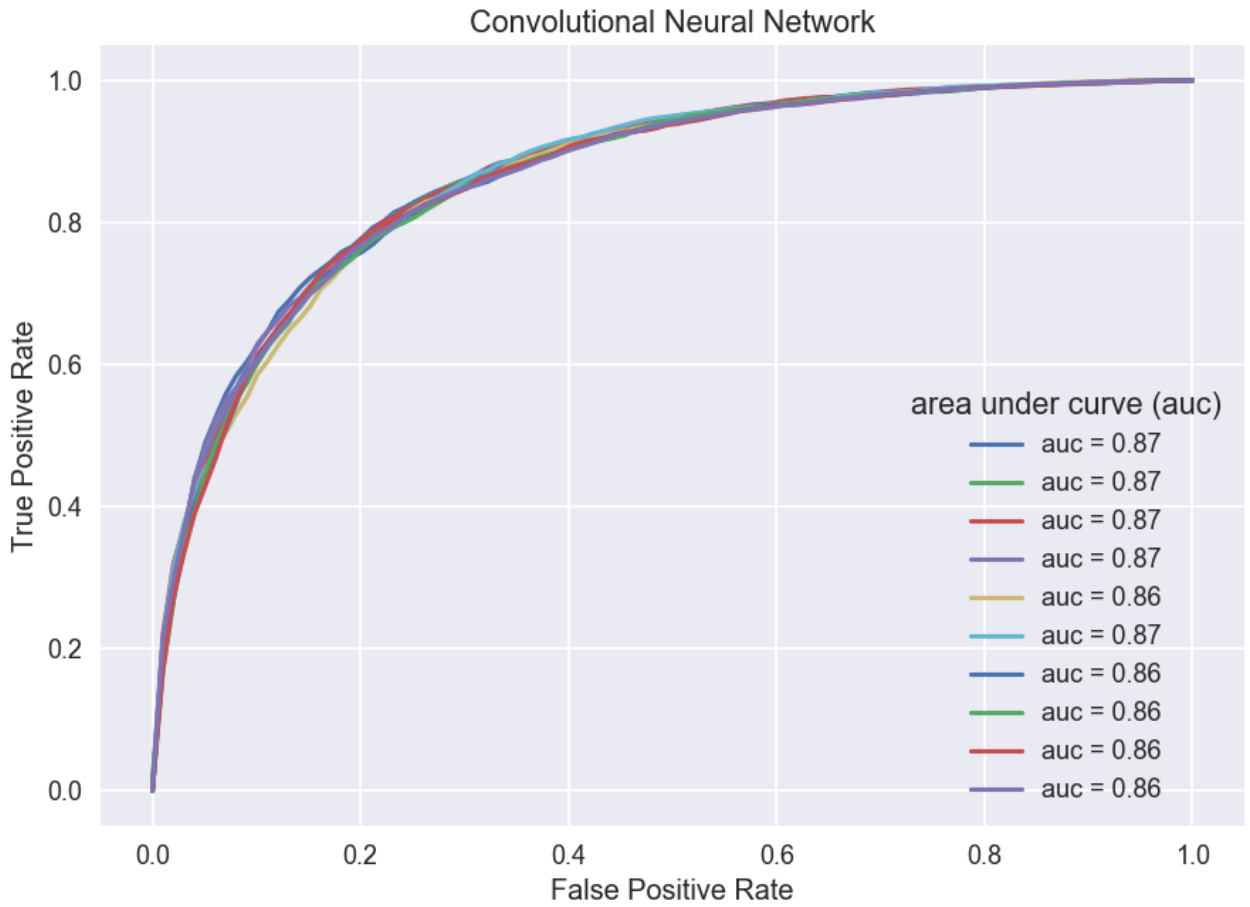
Advantage

- Best model among all
  - Higher scores

Disadvantage

- Computationally expensive

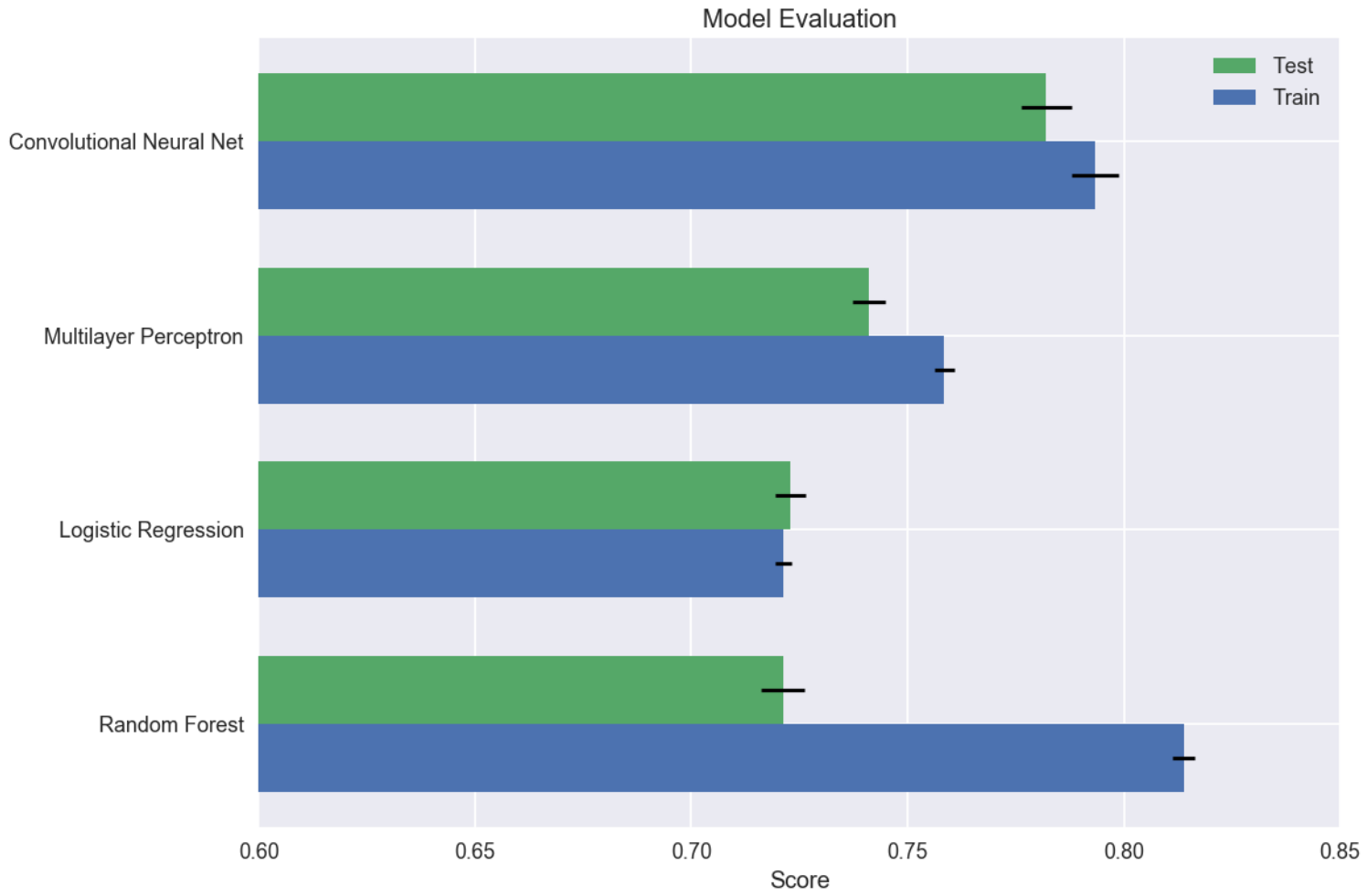
```
Using TensorFlow backend.

Loading signal data from data/signal_PU0_13TeV_MJ-65-95_PTJ-250-300.txt

Loading backgr data from data/backgr_PU0_13TeV_MJ-65-95_PTJ-250-300.txt

Processing 12679 signal and 12415 backgr samples

Training model

Training time = 1:37:54.676328
```



Model Evaluation

Training and Testing scores

Higher training score does not guarantee good performance (e.g. Random Forest overfits)



Machine Learning vs N-subjettiness

ML outperforms the traditional N-subjettiness approach