SM WZ diboson production in a three lepton final state

Data Analysis with ATLAS Open Data

Tiago Fernandes
Pedro Lagarelhos

Data Analysis in Particle Physics
Professor Nuno Castro

MAP JOINT DOCTORAL PROGRAMMES

Contents

Cut based analysis and dataset

Fiddling with the cuts

Machine Learning

Applying best models to test data

The (brief) CLs method

Conclusions and Future work

ATLAS Open Data

- Proton-proton (pp) collision data released by the ATLAS Collaboration to the public for educational purposes.
- Data collected by the ATLAS detector at the LHC at 13 TeV during the year 2016 and corresponding to an integrated luminosity of 10 fb^{-1}.
- Data is accompanied by a set of MC simulated samples describing several processes modeling the expected distributions of signal and background.

WZ production

- Production of WZ final states is an important test of the electroweak sector of the Standard Model since they arise from:
    - two vector bosons radiated by quarks or from the decay of a virtual W boson, which involves a triple gauge coupling;
    - vector-boson scattering processes, which involve triple and quartic gauge couplings and are sensitive to the electroweak symmetry breaking sector.
- New physics could manifest as a modification of these triple and quartic gauge coupling strength.
- The analysis illustrates the statistical limitations of the released dataset given the low production cross-section of the rare processes, where the variations between data and MC prediction are attributed to sizeable statistical fluctuations.

Dataset

- Data
- MC samples
  - WZ
  - WW, ZZ
  - V+jets
  - Top

- Event requirements
  - Exactly 3 leptons
  - Leading lepton p_T > 25 GeV

Object Pre-Selection Criteria

Electron	Muon
ID & ECAL reconstruction	ID & MS reconstruction
Loose identification	
Loose isolation	
$p_T > 7$ GeV	
$\lvert \eta \rvert < 2.47$	$\lvert \eta \rvert < 2.5$

Object Pre-Selection Criteria

Electron	**Muon**
Single electron trigger	Single muon trigger
Tight reconstruction	
$p_T > 20$ GeV	
Track cone isolation	
Calorimeter cone isolation	
$\lvert\eta\rvert < 2.47$ & $(\lvert\eta\rvert < 1.37$ & $\lvert\eta\rvert > 1.52)$	$\lvert\eta\rvert < 2.5$

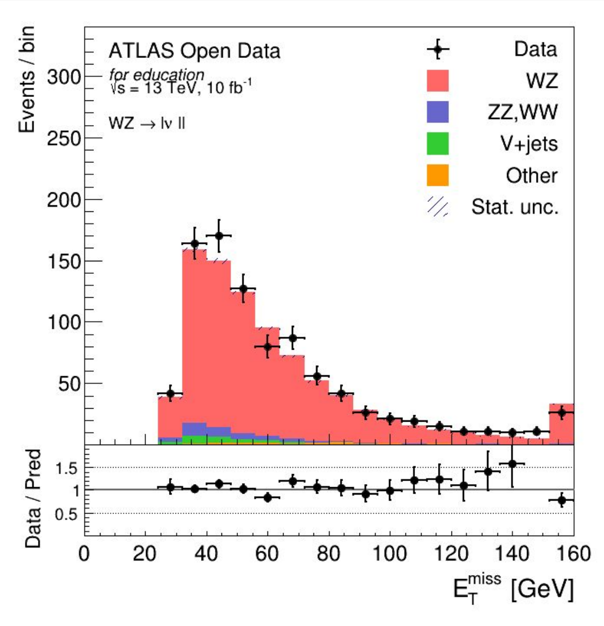
Event Selection Strategy

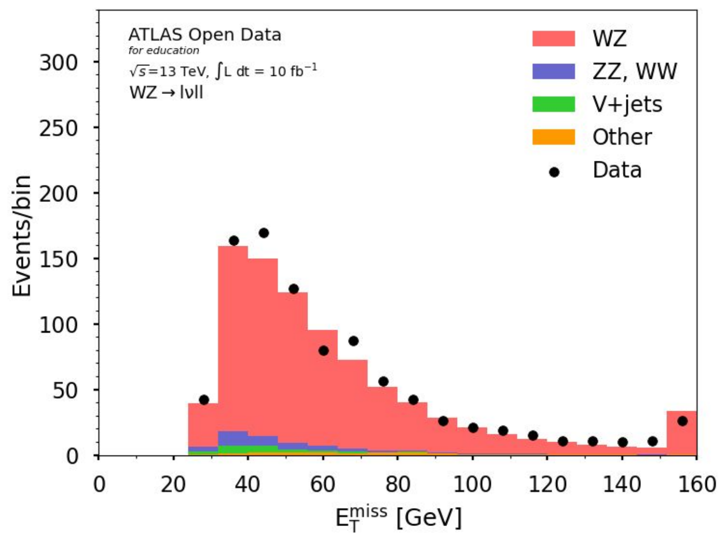
- Exactly 3 "good" leptons;
- 1 SFOS lepton pair;
- $| m_{ll} - m_Z | < 10$ GeV;
- $m_T^W > 30$ GeV;
- $E_T^{miss} > 30$ GeV;
- At least 1 lepton with $p_T > 25$ GeV.

Event Selection Strategy

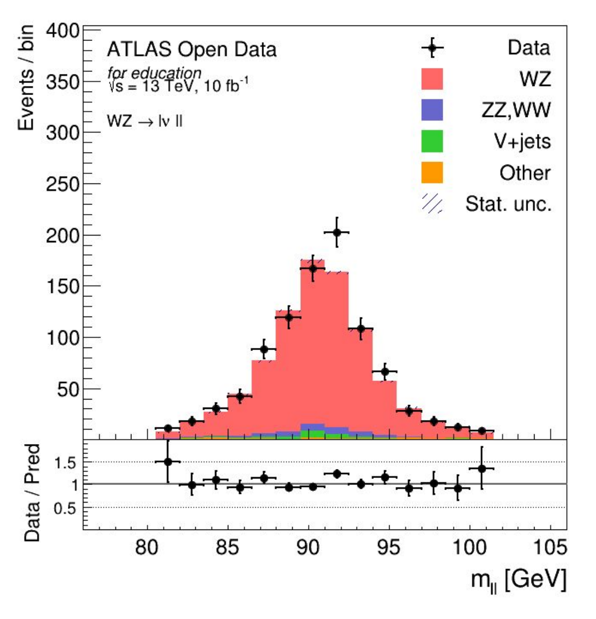
- Exactly 3 “good” leptons;
- 1 SFOS lepton pair;
- **$|m_{ll} - m_Z| < 10$ GeV;**
- **$m_T^W > 30$ GeV;**
- **$E_T^{miss} > 30$ GeV;**
- *At least 1 lepton with $p_T > 25$ GeV.*

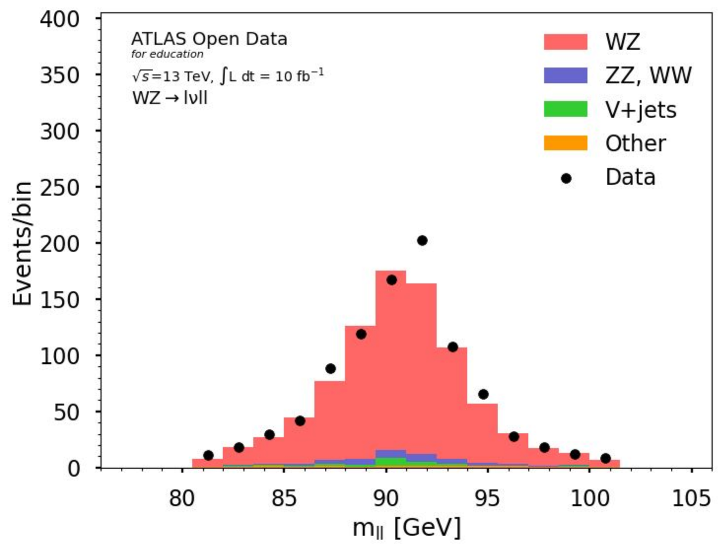
Recreating the analysis plots



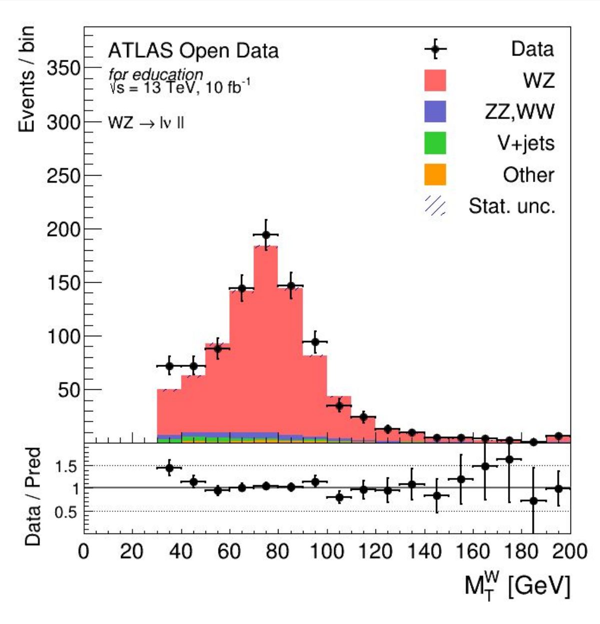


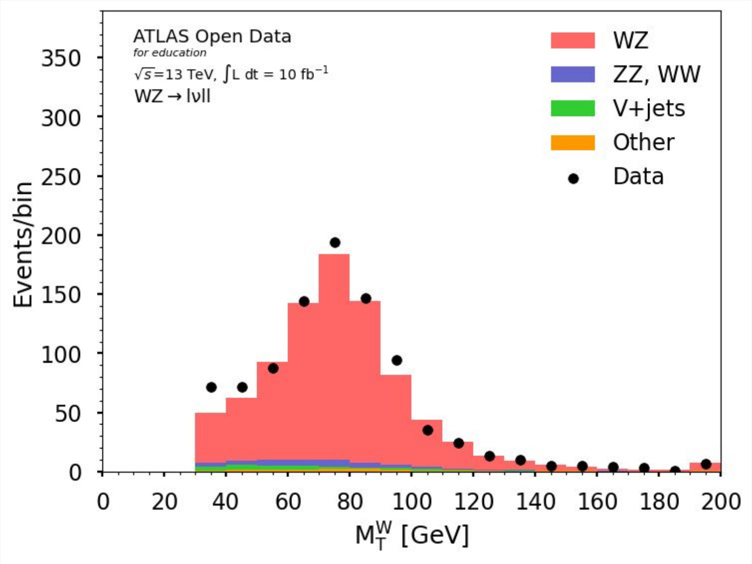
Recreating the analysis plots





Recreating the analysis plots

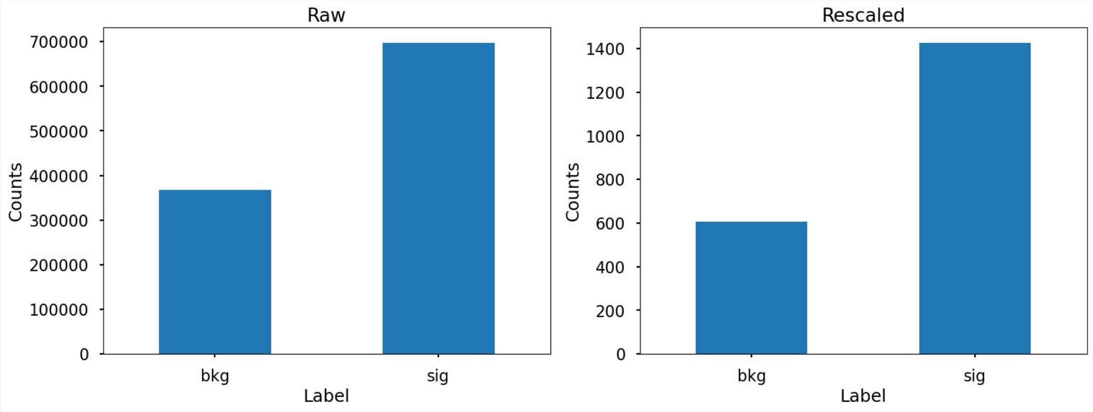




Exploring the cuts

- We chose the last cuts:
  - $|m_{ll} - m_Z| < 10$
  - $M_T^W > 30$
  - $E_T^{miss} > 30$
  - $p_T^{any\ lep} > 25$
    (has no effect)

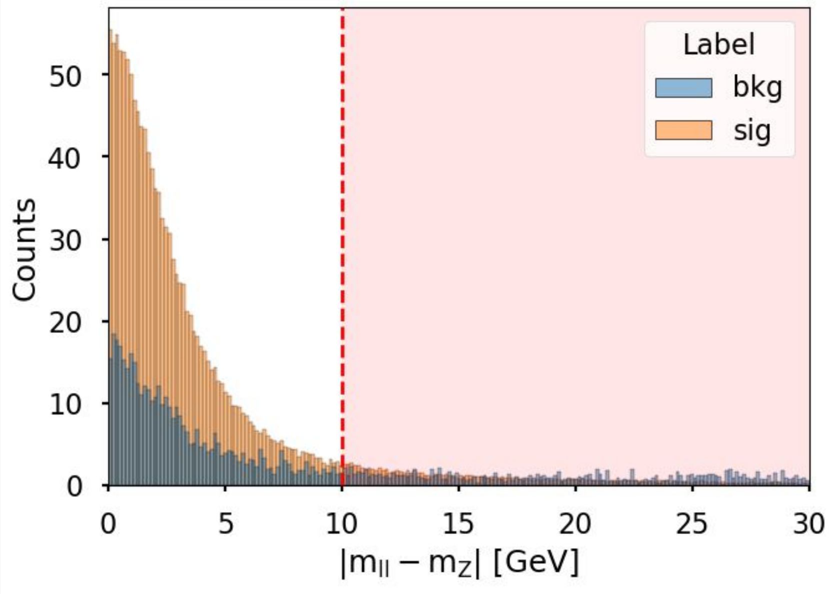
Number of events when the cuts are removed



signal - WZ, background - everything else

Exploring the cuts

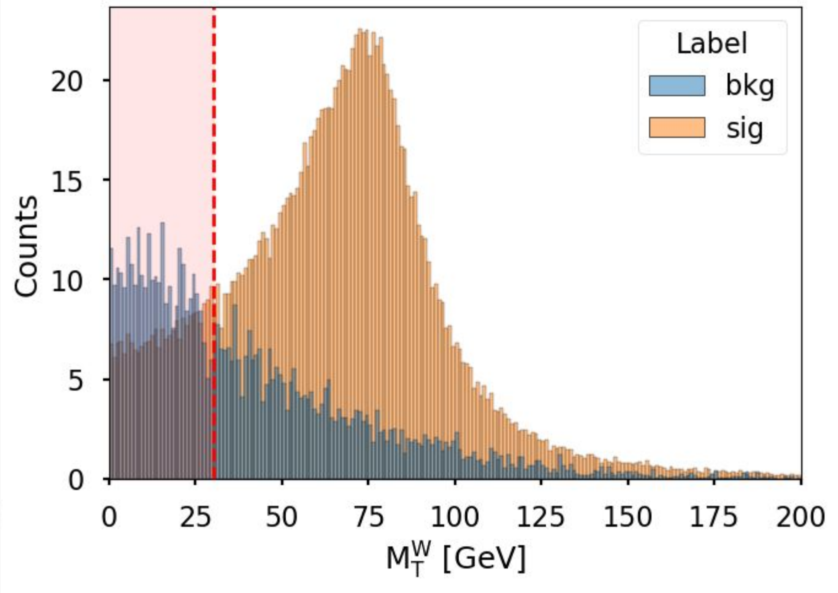
$$|\mathrm{m_{ll}} - \mathrm{m_Z}| < 10$$



- 10.93% signals cut
- 31.63% backgrounds cut

Exploring the cuts

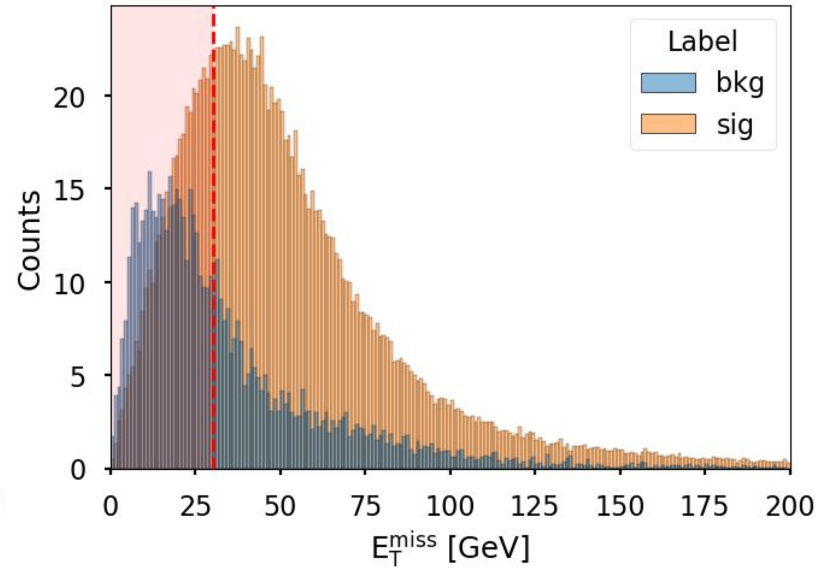
$M_T^W > 30$



- 15.52% signals cut
- 48.14% backgrounds cut

Exploring the cuts

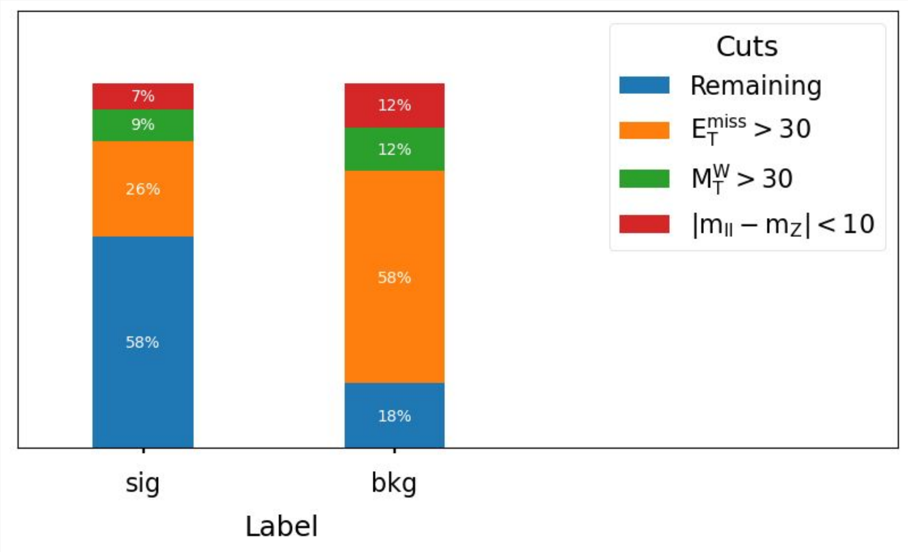
$$\mathrm{E_T^{miss}} > 30$$

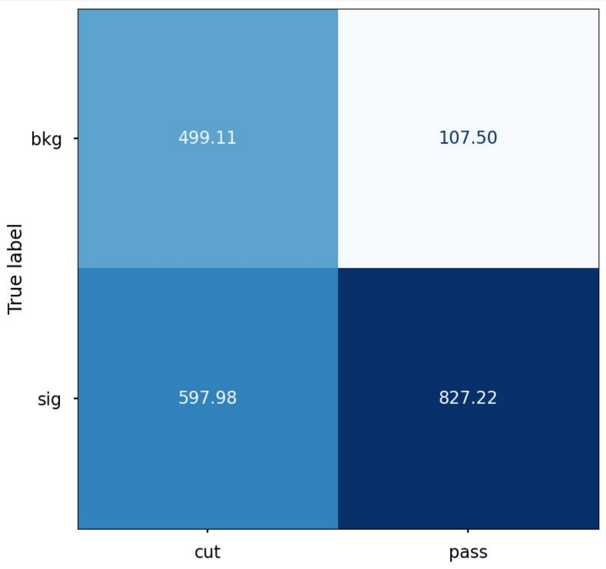


- 26.27% signals cut
- 58.35% backgrounds cut

Exploring the cuts

Combined effect of the cuts, ordered by most effective.





- #sig/#bkg = **7.69**
- **827** signals kept

ML
time

Approach

	lep1_ch	lep1_ID	lep2_ch	lep2_ID	lep3_ch	lep3_ID	mLL	zdiff	ptLL	etmiss	...
336596	1	2	2	2	1	2	96.502403	5.3148	374.562012	139.772003	...
1046443	1	2	1	2	2	2	89.442703	1.7449	43.378502	32.666199	...
994441	1	2	2	2	1	2	93.114700	1.9271	112.532997	123.005997	...
131389	2	1	2	1	1	1	87.930702	3.2569	155.957993	52.175999	...
45434	2	2	2	1	1	2	69.468399	21.7192	166.714005	68.529602	...
...	...	...	...	...	...	...	...	...	...	...	...
1006448	1	2	2	1	2	2	94.582497	3.3949	97.462898	5.116620	...
928535	2	2	1	2	2	2	90.003304	1.1843	28.004601	33.743801	...
489160	2	1	1	1	2	1	94.835403	3.6478	54.816700	29.547701	...
367127	1	2	2	2	1	1	92.213501	1.0259	133.578995	16.742399	...
350814	2	2	1	2	1	2	90.093002	1.0946	156.731995	81.692902	...

X

label
1
0
0
1
1
...
0
0
1
1
1

y

rescaled_weight
0.000622
0.000402
0.000627
0.000868
0.000367
...
0.000314
0.000115
0.009223
0.001042
0.001357

sample_weights

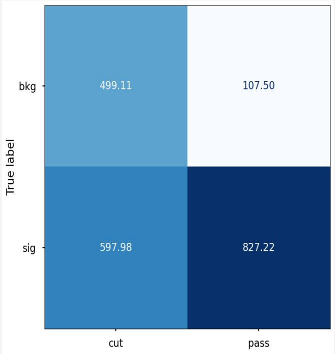
y == 1: WZ (sig)

y == 0: other processes (bkg)

- Tabular data -> tree-based models
  - Random Forest, Decision Trees, Gradient Boosted Trees
- Consider using sample weights to scale the importance of each row.
- Training on the Monte Carlo data (90% train + validation, 10% test).

Approach

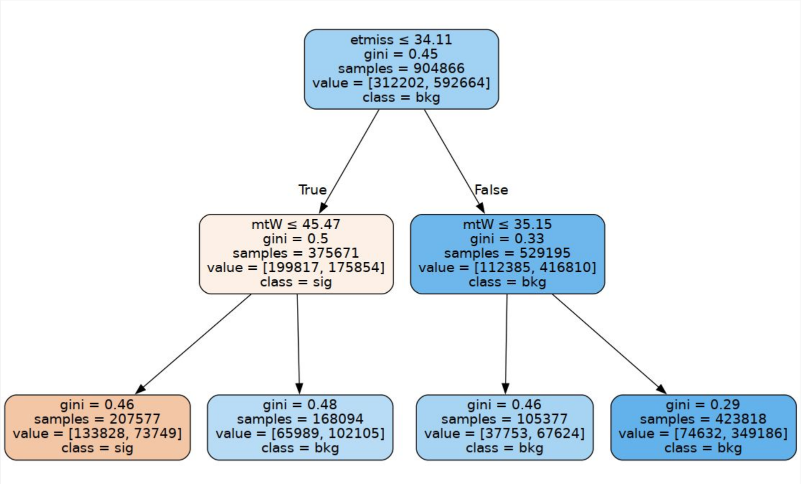
- Task: separate signal (WZ) from background (other processes)
- Goal:
  - obtain high ratio of signals to backgrounds...
  - while finding a good amount of signals.
- Metrics:
  - precision TP / (TP + FN);
  - percentage of true positives.



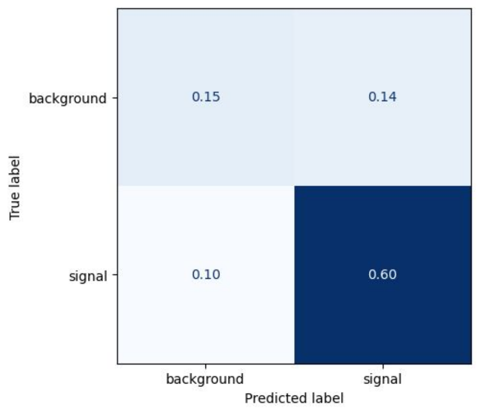
#sig/#bkg = **7.69** <-> precision: **0.885**

827 signals kept <-> **0.407** true positives

Decision Tree



- Very simple classifier, easily interpretable.
- Recovers cuts resembling the original analysis.



precision = **0.81**

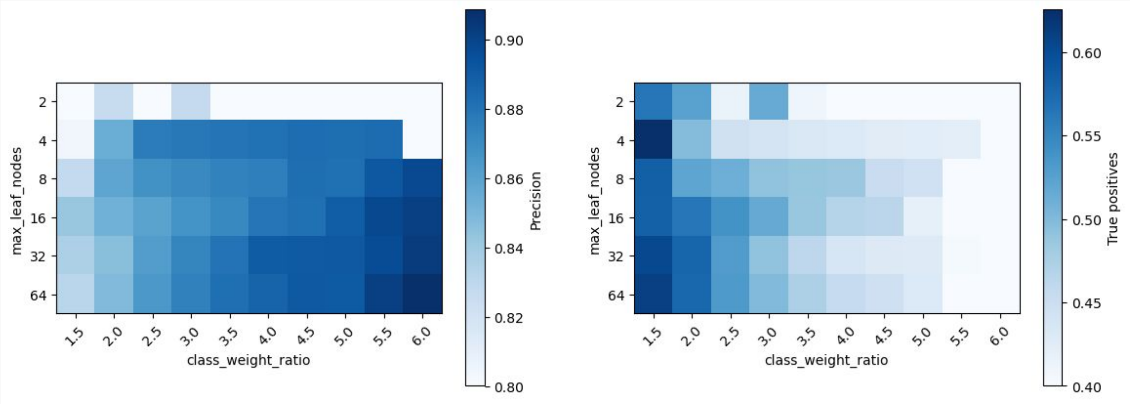
TPs = **0.60**

Better Decision Trees

- Grid searches to find optimal:
  - maximum number of leaf nodes
  - class_weight

… while using sample_weights during fit.

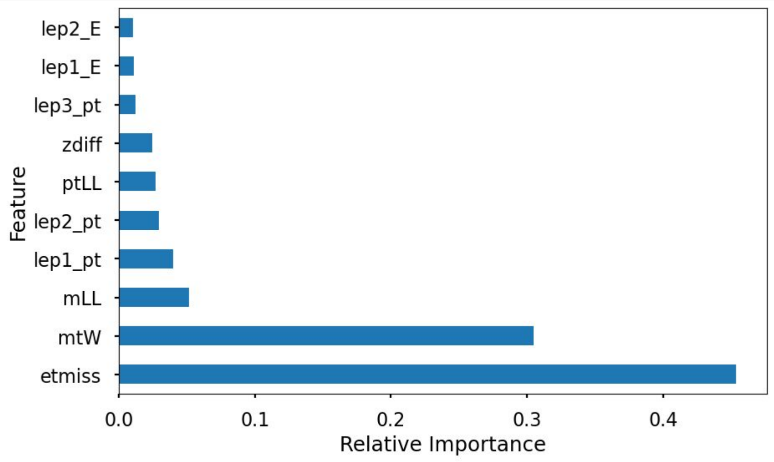
- Found model with max 64 leaf nodes and class weight ratio of 5.5 with:
  - **0.901** precision
  - **0.388** TPs



- Issue: negative sample weights result in probabilities $\notin [0, 1]$
- Next models do not use sample weights during fit.

Random Forests

- Random search for hyperparameter tuning of:
    - number of estimators
    - maximum leaf nodes
    - use of bootstrap
    - class weights
- Found a model with max 100 estimators, max 256 leaf nodes, no bootstrap, and class weight ratio of 2.5 with:
    - **0.874** precision
    - **0.408** TPs

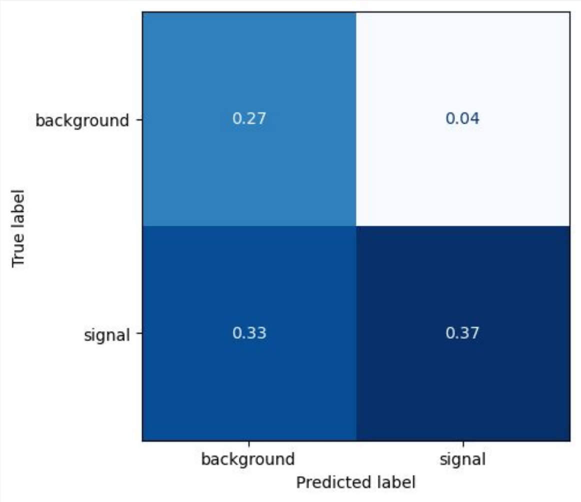


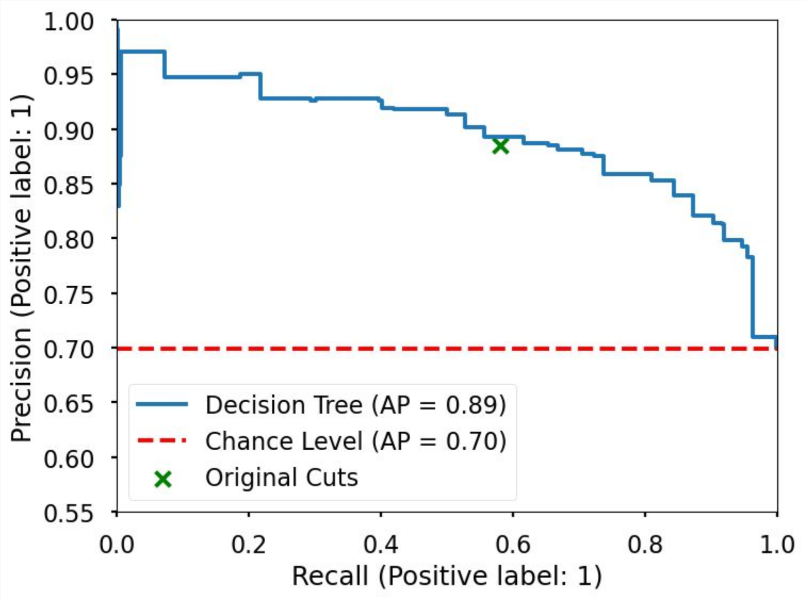
Boosted Decision Trees

- Random search for hyperparameter tuning of:
    - learning rate
    - maximum depth
    - minimum samples per leaf
    - l2 regularization
    - maximum number of bins
    - maximum number of iterations
    - class weights

- Found a model with lr=0.1, max depth of 3, min 20 samples per leaf, l2=0.001, max 127 bins, max 500 iterations and a class weight ratio of 2 with:
    - **0.865** precision
    - **0.483** TPs

DT evaluation on test set

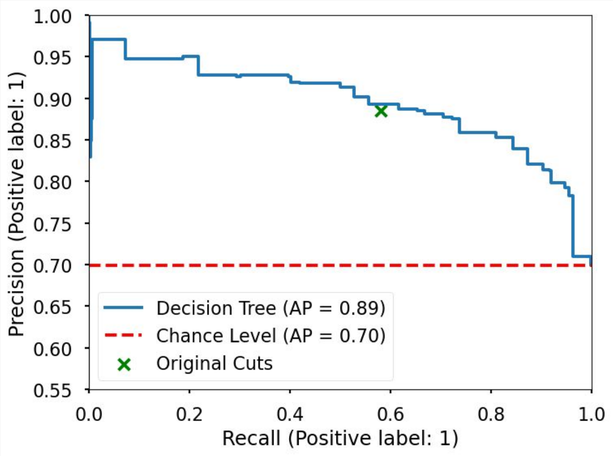


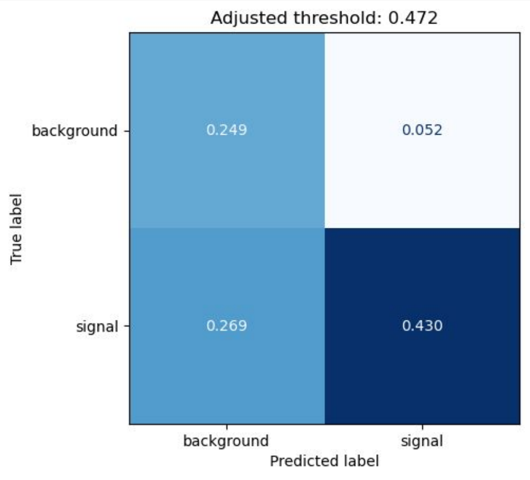


#sig/#bkg = **10.51** ✓

TPs = **0.369** ✗

DT evaluation on test set

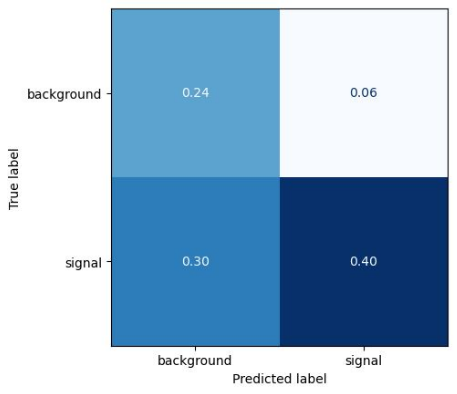


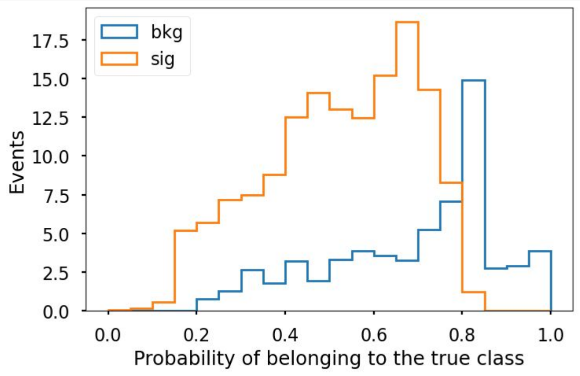


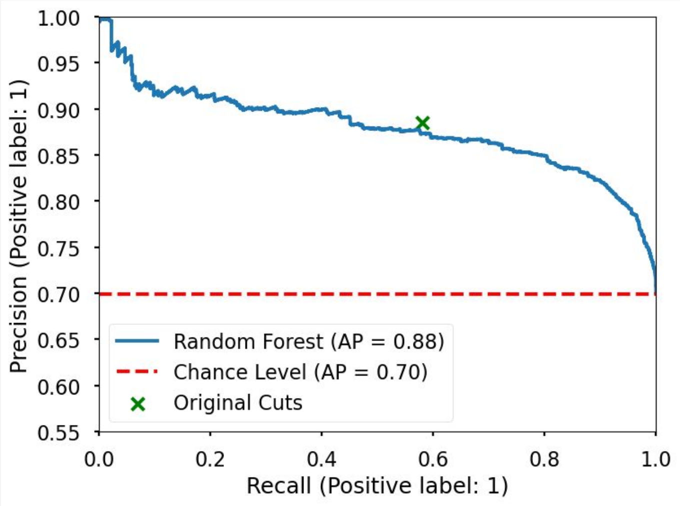
#sig/#bkg = **8.35** ✓

TPs = **0.430** ✓

RF evaluation on test set



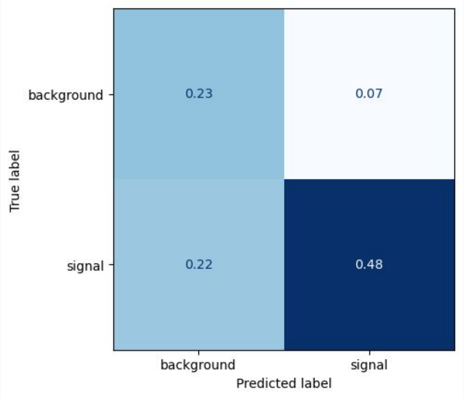


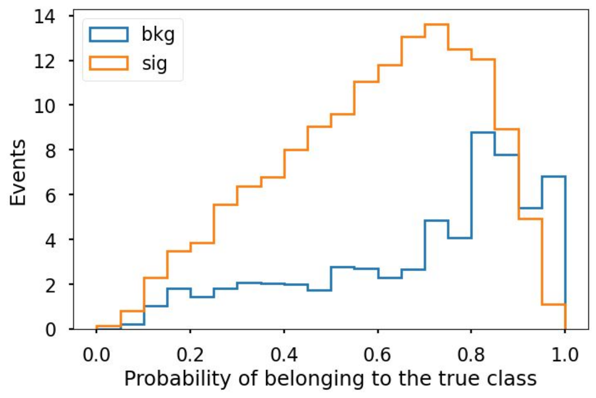


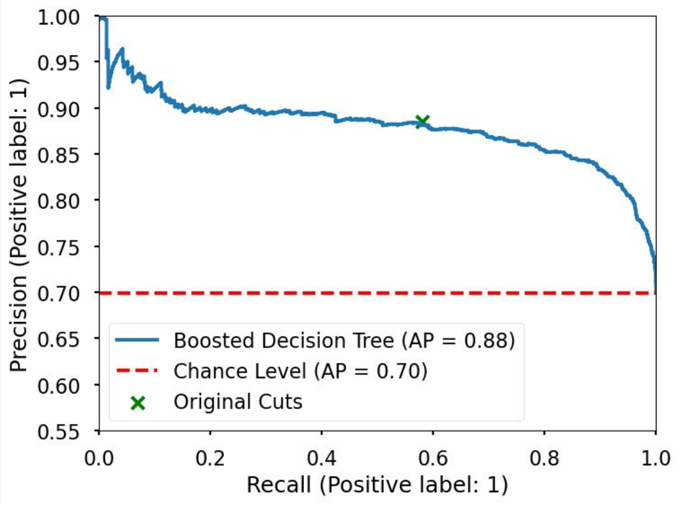
#sig/#bkg = **7.15**

TPs = **0.401**

BDT evaluation on test set



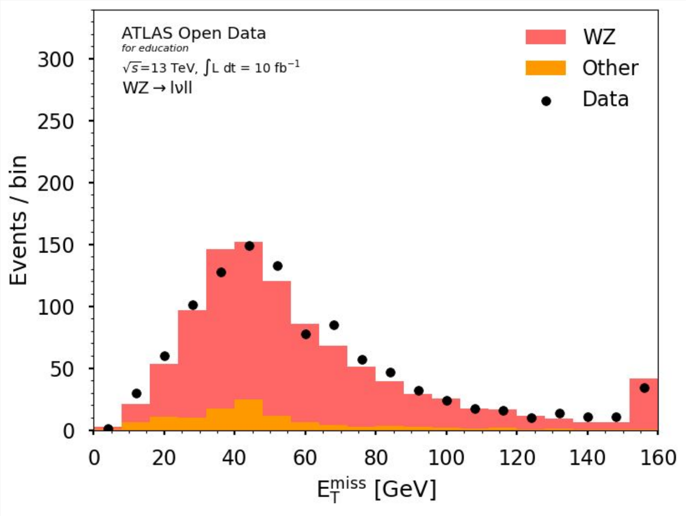




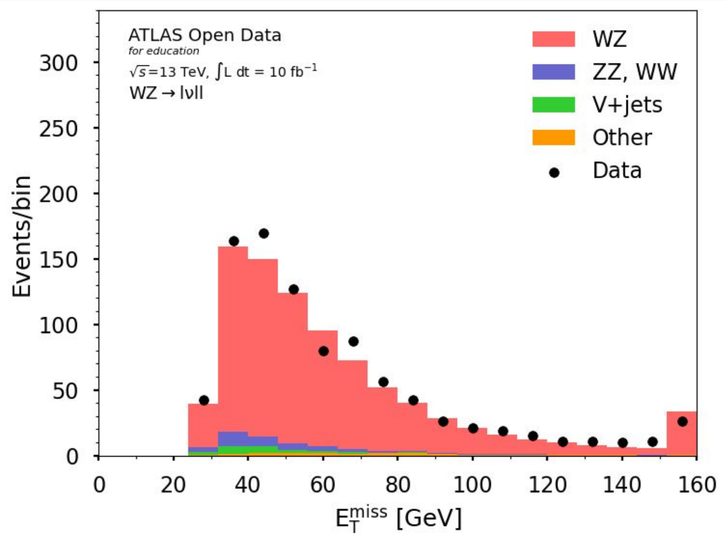
#sig/#bkg = **6.95**

TPs = **0.476**

Applying cuts to data



Our model
(Decision Tree)

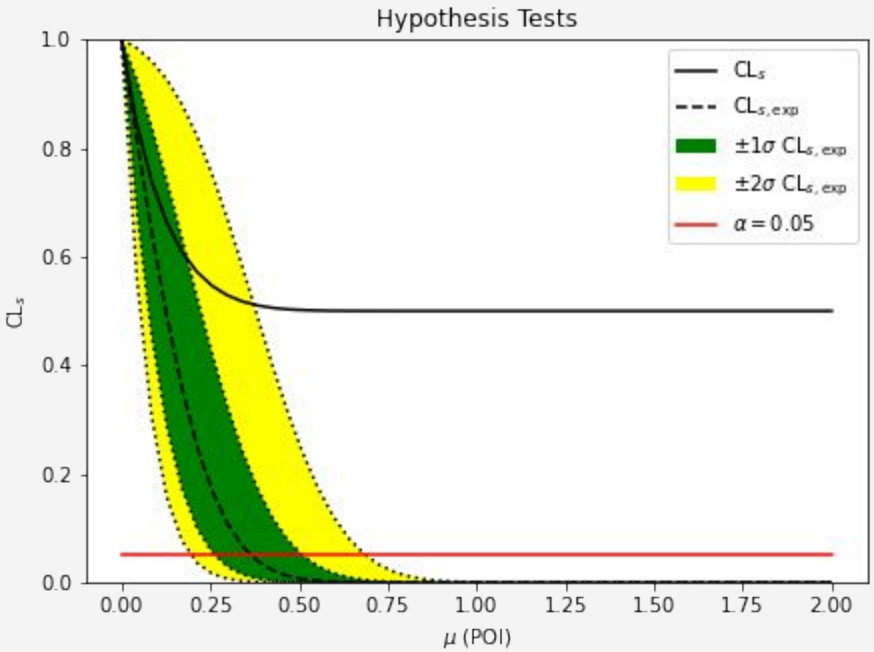


Original cuts

The CLs Method

CLs Method – Signal Strength

σ



$$\mu = \frac{\sigma_{95\%}}{\sigma_{th}}$$

	μ
-2σ	0.20
-1σ	0.27
Mean	0.37
1σ	0.51
2σ	0.69

Conclusions

- It was attempted to replace the final cuts of the analysis by a simple ML model.
- A Decision Tree was able to improve the analysis in terms of both number of signals and ratio of signals to backgrounds.
- More advanced tree methods were briefly explored, yielding results close to the analysis, but worse than the simple Decision Tree.
- We were able to recover a good amount of WZ signals with $E_T^{miss} < 30$ GeV, which would be thrown away by the classical analysis.

Further Work

- Relax more cuts, *e.g.* leave the object selection for an ML model instead of a standard cut based analysis;
- Understand how to better deal with the sample weighting.
- Apply an optimiser for ML hyperparameter optimization, *e.g.* Optuna;
- Further explore the CLs method and calculate limits on cross-sections;
- Explore other observables.

Thank you

for your attention