Testing the Efficiency of Machine Learning at Classifying Higgs Boson Decays and Testing the Efficiency of New Triggers in Run 3 data

Mentee: Parveen Narula (Beloit College)
Mentors: Cristina Mantilla Suarez (Fermilab), Raghav Kansal (UCSD & Fermilab)

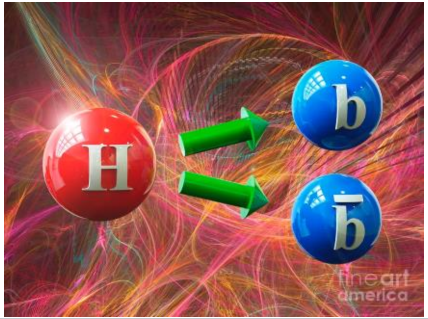
Introduction - Classifying Higgs Boson Decays

- LHC needs machine learning algorithms/Neural Networks to identify signal events vs QCD (background)

- Testing a machine learning (ML) algorithm on simulation data shows how well it can classify an event.

- Most common Higgs decay: H to bb

- My project used Monte Carlo simulation data of H to bb decay to test a ML algorithm.



Method: How the data I used was produced

- The simulated data has labels classifying each jet in each event as a H to bb decay event or QCD

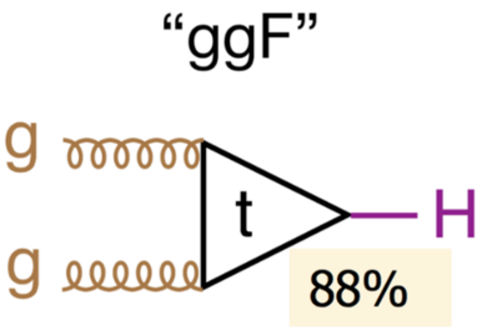
- The simulated data without the labels was run through a ML algorithm.

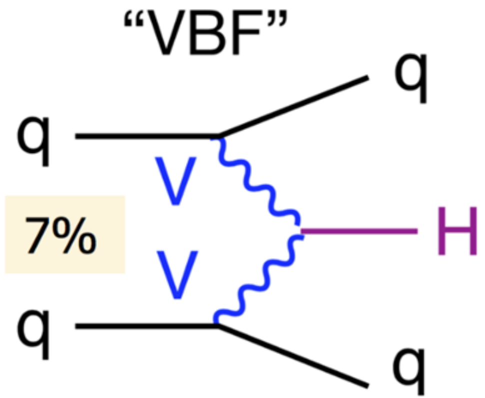
- For each jet the ML algorithm outputted a probability that the jet had a H to bb decay in it.

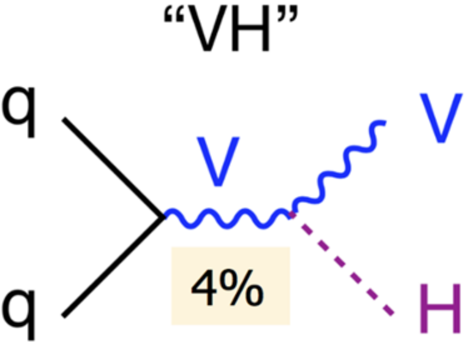
- Probability of jet containing H to bb decay: P(Xbb).

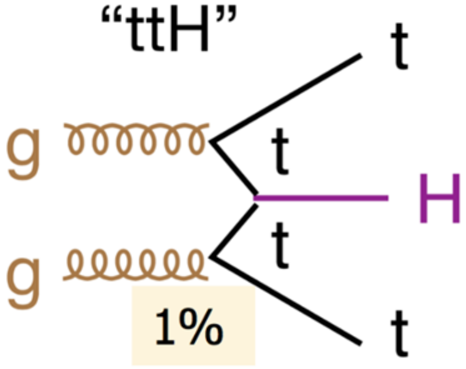
- Probability of a jet being QCD: P(QCD)

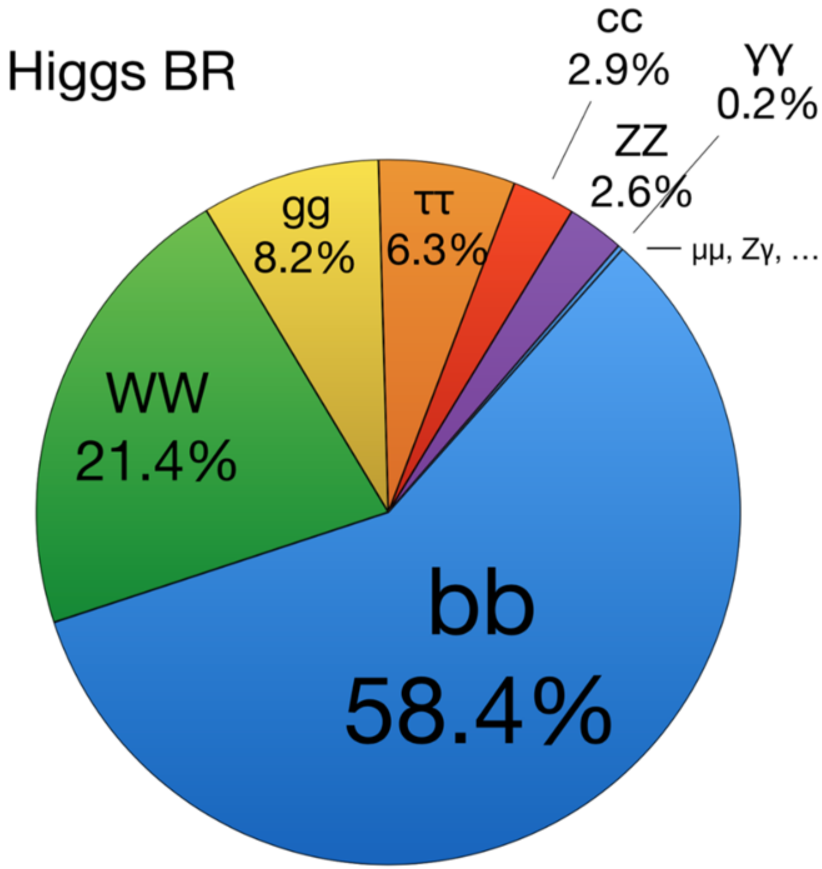
- P(Txbb) = P(Xbb) / (P(Xbb) + P(QCD))


"ggF"
g
g
t
H
88%


"VBF"
q
q
V
V
H
q
q
7%


"VH"
q
q
V
V
H
4%


"ttH"
g
g
t
t
t
t
H
1%


Higgs BR
cc
2.9%
γγ
0.2%
ZZ
2.6%
μμ, Zγ, ...
gg
8.2%
ττ
6.3%
WW
21.4%
bb
58.4%

Method: Organizing the data

- Given multiple QCD and Signal parquet files.
- File type (signal or QCD) corresponded with true classification of data.
- In my file for each event I was only given the jet with the highest Xbb.
- Each row corresponds to a jet and each column has particular information about the jet.
- Concatenated all the signal files into one pandas dataframe and all the QCD files into another. This is the signal dataframe.

	ak8FatJetEta	ak8FatJetPhi	ak8FatJetMass	ak8FatJetPt	ak8FatJetMsd	ak8FatJetParticleNetMD_QCD	ak8FatJetParticleNetMD_Xbb	ak8FatJetParticleNetMass	ak8FatJetParticleNetMD_Txbb	GenHiggsEta	GenHiggsPhi	GenHiggsMass	GenHiggsPt	GenHiggs_decay	ak8FatJetdRHqq	weight
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.316711	-1.669189	31.921875	251.125	6.687500	0.065125	0.934570	35.781250	0.934856	0.378906	-1.515625	125.0	243.0	1	0.165681	0.523666
1	-0.809937	0.872925	48.218750	250.250	9.804688	0.905273	0.000150	12.210938	0.000166	0.174805	-2.570312	125.0	193.5	1	3.005831	0.523666
2	1.638428	0.419678	31.468750	278.250	17.531250	0.988281	0.000034	5.171875	0.000034	0.906250	-2.835938	125.0	192.5	1	3.114846	0.523666
3	-0.023029	-1.786865	50.156250	264.500	3.177734	0.867676	0.000652	28.671875	0.000751	-0.489258	1.105469	125.0	223.0	1	2.929670	0.523666
4	-1.108154	-2.511230	38.250000	502.250	4.460938	0.668457	0.139038	17.812500	0.172184	-1.144531	-2.500000	125.0	473.0	1	0.038071	0.523666
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128675	0.901001	0.098938	123.375000	428.000	124.625000	0.005669	0.994141	134.125000	0.994330	0.910156	0.100830	125.0	454.0	1	0.009349	0.525195
128676	1.050537	-1.359619	57.000000	281.500	2.681641	0.166016	0.832520	54.906250	0.833741	1.042969	-1.542969	125.0	280.0	1	0.183506	0.525195
128677	0.651733	0.376709	124.125000	313.000	126.437500	0.008072	0.988281	131.375000	0.991899	0.671875	0.369141	125.0	321.0	1	0.021517	0.525195
128678	0.124741	-2.375488	61.312500	258.500	61.562500	0.233276	0.003538	54.500000	0.014941	0.087646	0.677734	125.0	242.0	1	3.053448	0.525195
128679	-1.190186	-1.547363	81.000000	300.750	81.375000	0.006447	0.954102	94.500000	0.993288	-1.195312	-1.632812	125.0	321.0	1	0.085603	0.525195

Method: Organizing the data

- I took the P(Txbb) column from the signal dataframe and put it in its own dataframe with a column name Probability.

- I added another column called Signal Marker where all the values were 1.

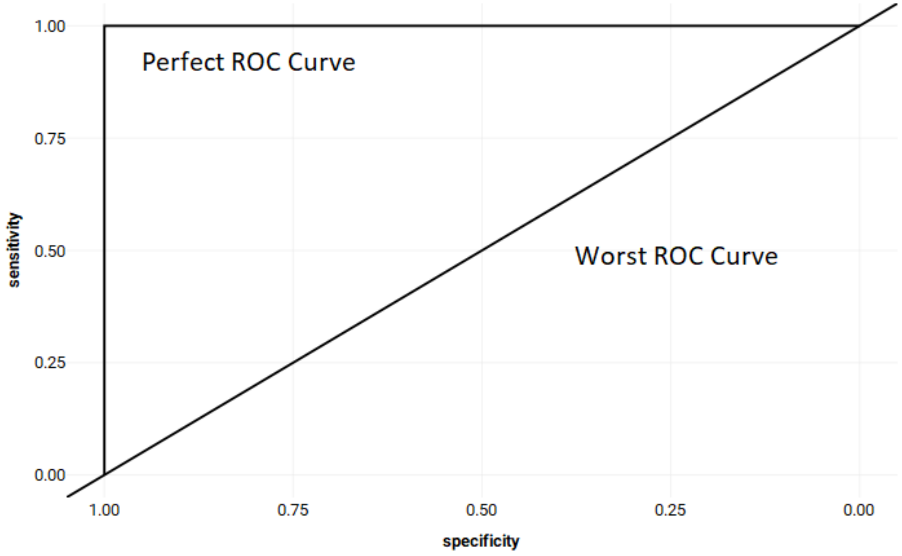
- I repeated this process for QCD data but made the column Signal Marker be filled with 0's .

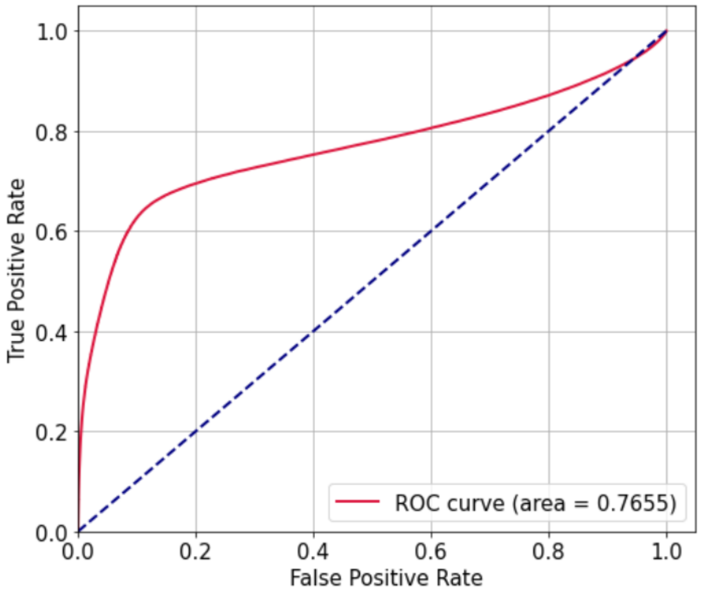
- I took these two data frames and concatenated them with each other to get one data frame. The signal marker column is there to identify whether that data came from a signal or QCD file.

	Probability	Signal Marker
0	0.934856	1
1	0.172184	1
2	0.983000	1
3	0.464225	1
4	0.559822	1
...	...	...
3708415	0.763717	0
3708416	0.000772	0
3708417	0.003312	0
3708418	0.052866	0
3708419	0.006005	0

Method - Making a ROC curve

- I made a roc curve using the probability dataframe to show how effective our ML algorithm was at identifying H to bb deacy events.



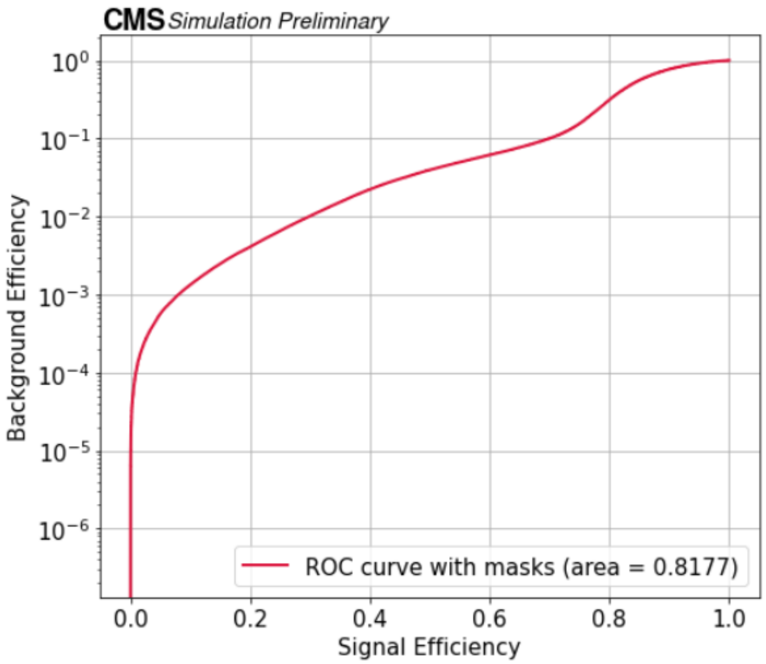


Method - Selections to get a better result

- I applied multiple selections on my dataframe to only select certain jets.

- Selections used:
  - pt > 200
  - Distance between Jet and Higgs Boson < 0.8 (only applied to signal since this is a column only signal dataframe has)

- Different way to make a ROC curve



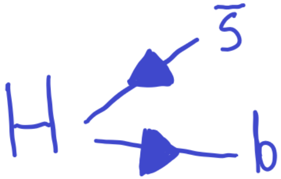
Conclusion and Future Work

Conclusion:

- Classifier is performing well but not at the level we expected

Future Work: H to bs decay

- Can use this technique to estimate if H to bs decay is sensitive enough to warrant looking for it in real LHC data.
- H to bs decay does not exist according to the Standard Model because of flavour violation.
- Some other theories predict H to bs decay
- This is beyond standard model research because if we find H to bs dec we know the standard model in its current state is incomplete or wrong.



Introduction - Triggers

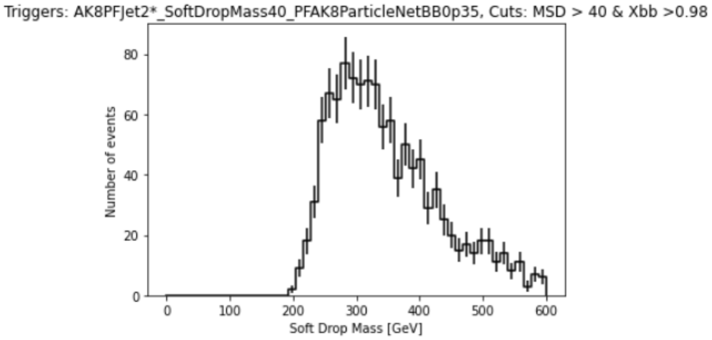
- A trigger is the program that makes split second decisions on what to keep and what do discard while the LHC is running.

- LHC doesn't have the storage capacity to keep all the data it is creating so a trigger is necessary.

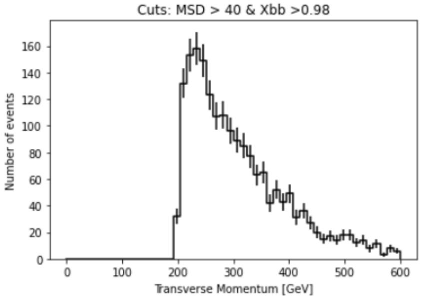
- I tested efficiency of the triggers currently being used in run 3 on real run 3 data to see if these triggers were more efficient than the triggers used in run 2.

Method - Testing Efficiency vs Pt

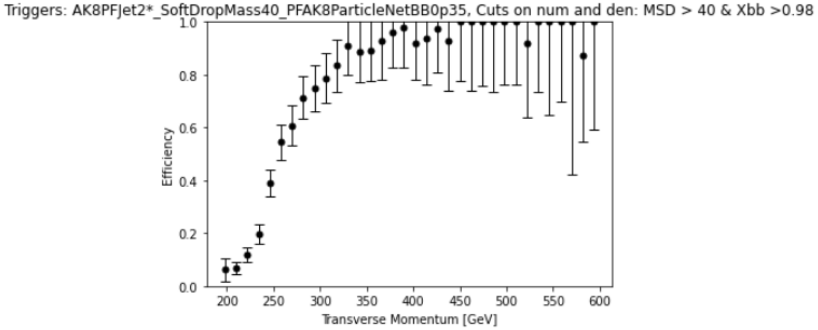
Hist with triggers:



Hist without triggers:

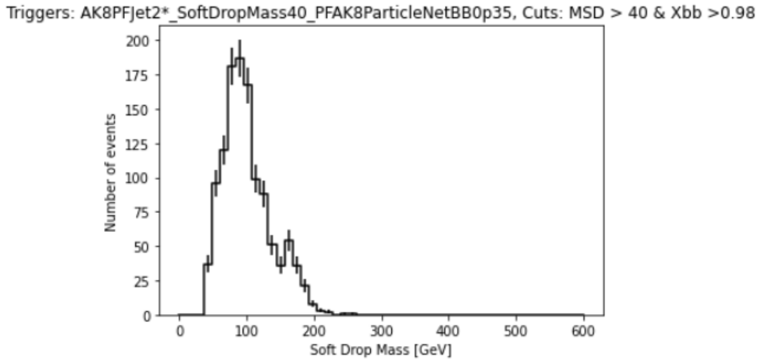


- Trigger efficiency found by: Hist with triggers applied/Hist without Triggers

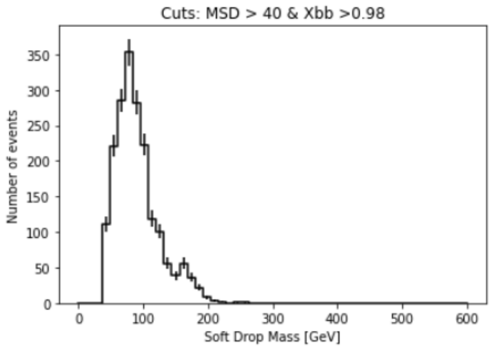


Method - Testing Efficiency vs Soft Drop Mass

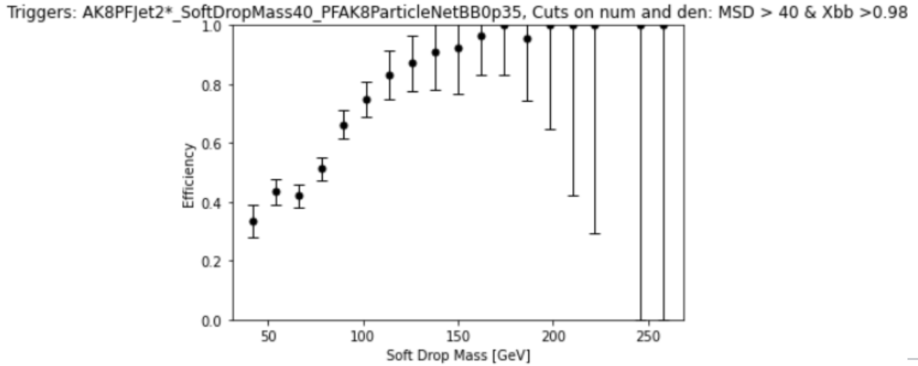
Hist with triggers:



Hist without triggers:

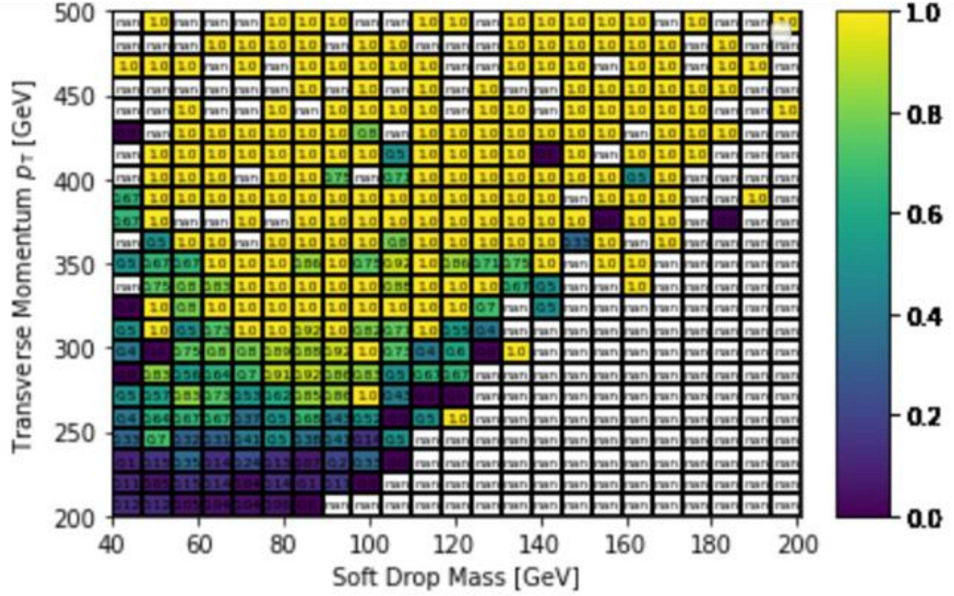


- Trigger efficiency found by: Hist with triggers applied/Hist without Triggers

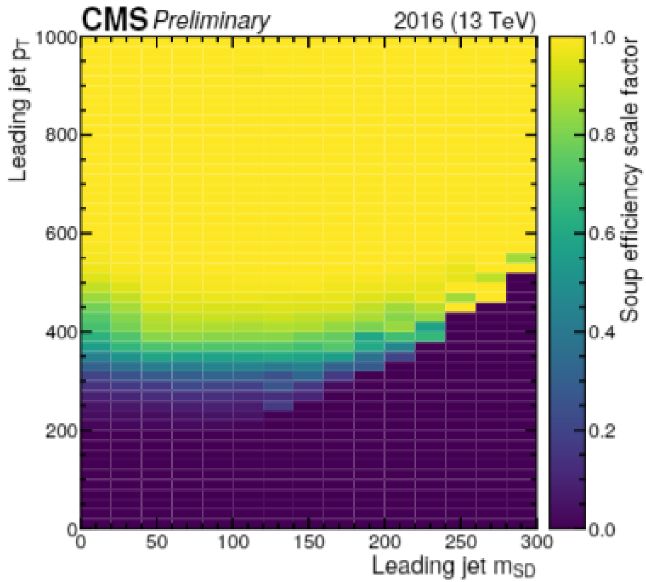


Comparing 2D plots of Run 2 and Run 3 Triggers

Run 3 triggers:



Run 2 Triggers:



- Run 3 triggers have close to 100% efficiency starting at pt ~300 GeV but Run 2 ~400 GeV.

Conclusion and Future Work

Conclusion:

- Run 3 triggers are performing very well, improving the efficiency for pt 300-400 GeV jets

Future Work:

- Compare overall sensitivity with new triggers vs Run 2 triggers

- Create more efficient triggers to use in future runs

THANK YOU