[ИИ, команда 67] EDA 1, Higgs Boson dataset

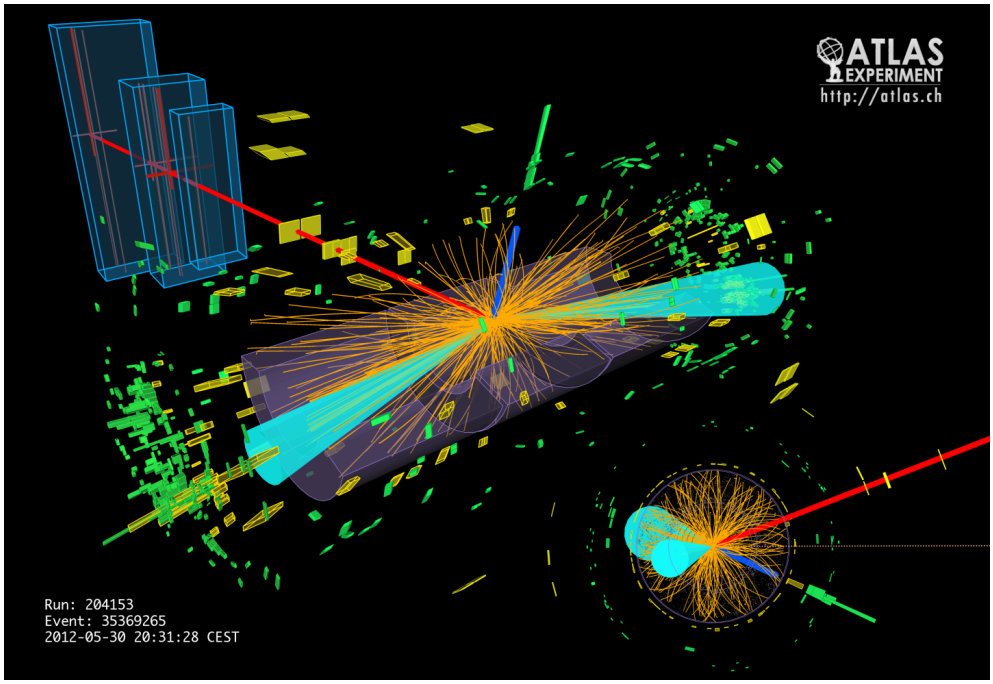


Fig 1. Higgs into fermions: Evidence of the Higgs boson decaying to fermions (image credit: CERN)

Ниже представлен разведочный анализ данных по датасету Higgs boson machine learning challenge собранному в ходе экспериментов ATLAS на большом адронном коллайдере (Large Hadron Collider - LHC).

Краткое описание признаков

1. **EventId** - Уникальный целочисленный идентификатор события.
2. **DER_mass_MMC** - Оценочная масса mH кандидата в бозоны Хиггса, полученная путем вероятностной интеграции фазового пространства.
3. **DER_mass_transverse_met_lep** - Поперечная масса (21) между недостающей поперечной энергией и лептоном.
4. **DER_mass_vis** - Инвариантная масса (20) адронного тау и лептона.
5. **DER_pt_h** - Модуль (19) векторной суммы поперечного импульса адронного тау, лептона и недостающего вектора поперечной энергии.
6. **DER_deltaeta_jet_jet** - Абсолютное значение псевдобыстротного разделения (22) между двумя струями (не определено, если PRI_jet_num ≤ 1).

7. **DER_mass_jet_jet** - Инвариантная масса (20) двух струй (не определена, если PRI_jet_num ≤ 1).
8. **DER_prodeta_jet_jet** - Произведение псевдобыстрот двух струй (не определена, если PRI_jet_num ≤ 1).
9. **DER_deltar_tau_lep** - Разделение R (23) между адронным тау и лептоном.
10. **DER_pt_tot** - Модуль (19) векторной суммы недостающих поперечных импульсов и поперечных импульсов адронного тау, лептона, ведущей струи (если PRI_jet_num ≥ 1) и следующей за ней струи (если PRI_jet_num = 2) (но не каких-либо дополнительных струй).
11. **DER_sum_pt** - Сумма модулей (19) поперечных импульсов адронного тау, лептона, ведущей струи (если PRI_jet_num ≥ 1) и следующей за ней струи (если PRI_jet_num = 2) и других струй (если PRI_jet_num = 3).
12. **DER_pt_ratio_lep_tau** - Отношение поперечных импульсов лептона и адронного тау.
13. **DER_met_phi_centrality** - Центральность азимутального угла недостающего поперечного вектора энергии.
14. **PRI_tau_pt** - Поперечный импульс адронного тау
15. **PRI_tau_eta** - Псевдобыстрота η адронного тау.
16. **PRI_tau_phi** - Азимутальный угол φ адронного тау.
17. **PRI_lep_pt** - Поперечный импульс лептона (электрона или мюона).
18. **PRI_lep_eta** - Псевдобыстрота η лептона.
19. **PRI_lep_phi** - Азимутальный угол φ лептона.
20. **PRI_met** - Недостающая поперечная энергия
21. **PRI_met_phi** - Азимутальный угол φ недостающей поперечной энергии
22. **PRI_met_sumet** - Полная поперечная энергия в детекторе.
23. **PRI_jet_num** - Количество струй (целое число со значением 0, 1, 2 или 3; возможные большие значения ограничены 3).
24. **PRI_jet_leading_pt** - Поперечный импульс ведущей струи, то есть струи с наибольшим поперечным импульсом (не определено, если PRI_jet_num = 0).
25. **PRI_jet_leading_eta** - Псевдобыстрота η ведущей струи
26. **PRI_jet_leading_phi** - Азимутальный угол φ ведущей струи
27. **PRI_jet_subleading_pt** - Поперечный импульс ведущей струи, то есть струи со вторым по величине поперечным импульсом
28. **PRI_jet_subleading_eta** - Псевдобыстрота η струи, идущей ниже
29. **PRI_jet_subleading_phi** - Азимутальный угол φ струи, идущей снизу
30. **PRI_jet_all_pt** - Скалярная сумма поперечного импульса всех струй событий.
31. **PRI_jet_all_pt** -
32. **Weight** - Вес события описан в разделе 3.3. Не использовать как признак. Недоступно в тестовом образце.

33. **Label** - Метка события y_i ∈ {s, b} (s - сигнальные события, b - фоновые события). Не использовать как признак. Недоступно в тестовом образце.

Фоновые события, сигнальные события и область выбора.

Фоновые события в основном производятся распадом частиц, которые, хотя и экзотические по своей природе, известны заранее из предыдущих поколений экспериментов. Целью анализа является поиск области (называемой областью выбора) в пространстве признаков, которая производит значительное превышение событий (называемых сигнальными событиями) по сравнению с тем, что могут объяснить известные фоновые процессы. После того, как область была зафиксирована, применяется статистический тест для определения значимости превышения. Если вероятность того, что превышение было произведено фоновыми процессами, падает ниже определенного предела, это указывает на обнаружение новой частицы.

Общая цель — улучшить процедуру, которая создает область выбора, т. е. область в пространстве признаков, которая производит события сигнала.

EDA

Сравним тренировочный и тестовый наборы:

	Тренировочный набор	Тестовый набор
Количество событий	250000	550000
Количество дублей	0	0
Количество пропущенных значений	0	0

Типы данных:

```
EventId                     int64
DER_mass_MMC                float64
DER_mass_transverse_met_lep float64
DER_mass_vis                float64
```

DER_pt_h	float64
DER_deltaeta_jet_jet	float64
DER_mass_jet_jet	float64
DER_prodeta_jet_jet	float64
DER_deltar_tau_lep	float64
DER_pt_tot	float64
DER_sum_pt	float64
DER_pt_ratio_lep_tau	float64
DER_met_phi_centrality	float64
DER_lep_eta_centrality	float64
PRI_tau_pt	float64
PRI_tau_eta	float64
PRI_tau_phi	float64
PRI_lep_pt	float64
PRI_lep_eta	float64
PRI_lep_phi	float64
PRI_met	float64
PRI_met_phi	float64
PRI_met_sumet	float64
PRI_jet_num	int64
PRI_jet_leading_pt	float64
PRI_jet_leading_eta	float64
PRI_jet_leading_phi	float64
PRI_jet_subleading_pt	float64
PRI_jet_subleading_eta	float64
PRI_jet_subleading_phi	float64

```
PRI_jet_all_pt                  float64

Weight                          float64

Label                            object
```

Описательные характеристики данных:

Name: EventId, count: 250000.000000, mean: 224999.500000, std: 72168.927986, min: 100000.000000, 25%: 162499.750000, 50%: 224999.500000, 75%: 287499.250000, max: 349999.000000;

Name: DER_mass_MMC, count: 250000.000000, mean: -49.023079, std: 406.345647, min: -999.000000, 25%: 78.100750, 50%: 105.012000, 75%: 130.606250, max: 1192.026000;

Name: DER_mass_transverse_met_lep, count: 250000.000000, mean: 49.239819, std: 35.344886, min: 0.000000, 25%: 19.241000, 50%: 46.524000, 75%: 73.598000, max: 690.075000;

Name: DER_mass_vis, count: 250000.000000, mean: 81.181982, std: 40.828691, min: 6.329000, 25%: 59.388750, 50%: 73.752000, 75%: 92.259000, max: 1349.351000;

Name: DER_pt_h, count: 250000.000000, mean: 57.895962, std: 63.655682, min: 0.000000, 25%: 14.068750, 50%: 38.467500, 75%: 79.169000, max: 2834.999000;

Name: DER_deltaeta_jet_jet, count: 250000.000000, mean: -708.420675, std: 454.480565, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: 0.490000, max: 8.503000;

Name: DER_mass_jet_jet, count: 250000.000000, mean: -601.237051, std: 657.972302, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: 83.446000, max: 4974.979000;

Name: DER_prodeta_jet_jet, count: 250000.000000, mean: -709.356603, std: 453.019877, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: -4.593000, max: 16.690000;

Name: DER_deltar_tau_lep, count: 250000.000000, mean: 2.373100, std: 0.782911, min: 0.208000, 25%: 1.810000, 50%: 2.491500, 75%: 2.961000, max: 5.684000;

Name: DER_pt_tot, count: 250000.000000, mean: 18.917332, std: 22.273494, min: 0.000000, 25%: 2.841000, 50%: 12.315500, 75%: 27.591000, max: 2834.999000;

Name: DER_sum_pt, count: 250000.000000, mean: 158.432217, std: 115.706115, min: 46.104000, 25%: 77.550000, 50%: 120.664500, 75%: 200.478250, max: 1852.462000;

Name: DER_pt_ratio_lep_tau, count: 250000.000000, mean: 1.437609, std: 0.844743, min: 0.047000, 25%: 0.883000, 50%: 1.280000, 75%: 1.777000, max: 19.773000;

Name: DER_met_phi_centrality, count: 250000.000000, mean: -0.128305, std: 1.193585, min: -1.414000, 25%: -1.371000, 50%: -0.356000, 75%: 1.225000, max: 1.414000;

Name: DER_lep_eta_centrality, count: 250000.000000, mean: -708.985189, std: 453.596721, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: 0.000000, max: 1.000000;

Name: PRI_tau_pt, count: 250000.000000, mean: 38.707419, std: 22.412081, min: 20.000000, 25%: 24.591750, 50%: 31.804000, 75%: 45.017000, max: 764.408000;

Name: PRI_tau_eta, count: 250000.000000, mean: -0.010973, std: 1.214079, min: -2.499000, 25%: -0.925000, 50%: -0.023000, 75%: 0.898000, max: 2.497000;

Name: PRI_tau_phi, count: 250000.000000, mean: -0.008171, std: 1.816763, min: -3.142000, 25%: -1.575000, 50%: -0.033000, 75%: 1.565000, max: 3.142000;

Name: PRI_lep_pt, count: 250000.000000, mean: 46.660207, std: 22.064922, min: 26.000000, 25%: 32.375000, 50%: 40.516000, 75%: 53.390000, max: 560.271000;

Name: PRI_lep_eta, count: 250000.000000, mean: -0.019507, std: 1.264982, min: -2.505000, 25%: -1.014000, 50%: -0.045000, 75%: 0.959000, max: 2.503000,

Name: PRI_lep_phi, count: 250000.000000, mean: 0.043543, std: 1.816611, min: -3.142000, 25%: -1.522000, 50%: 0.086000, 75%: 1.618000, max: 3.142000;

Name: PRI_met, count: 250000.000000, mean: 41.717235, std: 32.894693, min: 0.109000, 25%: 21.398000, 50%: 34.802000, 75%: 51.895000, max: 2842.617000;

Name: PRI_met_phi, count: 250000.000000, mean: -0.010119, std: 1.812223, min: -3.142000, 25%: -1.575000, 50%: -0.024000, 75%: 1.561000, max: 3.142000;

Name: PRI_met_sumet, count: 250000.000000, mean: 209.797178, std: 126.499506, min: 13.678000, 25%: 123.017500, 50%: 179.739000, 75%: 263.379250, max: 2003.976000;

Name: PRI_jet_num, count: 250000.000000, mean: 0.979176, std: 0.977426, min: 0.000000, 25%: 0.000000, 50%: 1.000000, 75%: 2.000000, max: 3.000000;

Name: PRI_jet_leading_pt, count: 250000.000000, mean: -348.329567, std: 532.962789, min: -999.000000, 25%: -999.000000, 50%: 38.960000, 75%: 75.349000, max: 1120.573000;

Name: PRI_jet_leading_eta, count: 250000.000000, mean: -399.254314, std: 489.338286, min: -999.000000, 25%: -999.000000, 50%: -1.872000, 75%: 0.433000, max: 4.499000;

Name: PRI_jet_leading_phi, count: 250000.000000, mean: -399.259788, std: 489.333883, min: -999.000000, 25%: -999.000000, 50%: -2.093000, 75%: 0.503000, max: 3.141000;

Name: PRI_jet_subleading_pt, count: 250000.000000, mean: -692.381204, std: 479.875496, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: 33.703000, max: 721.456000;

Name: PRI_jet_subleading_eta, count: 250000.000000, mean: -709.121609, std: 453.384624, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: -2.457000, max: 4.500000;

Name: PRI_jet_subleading_phi, count: 250000.000000, mean: -709.118631, std: 453.389017, min: -999.000000, 25%: -999.000000, 50%: -999.000000, 75%: -2.275000, max: 3.142000,

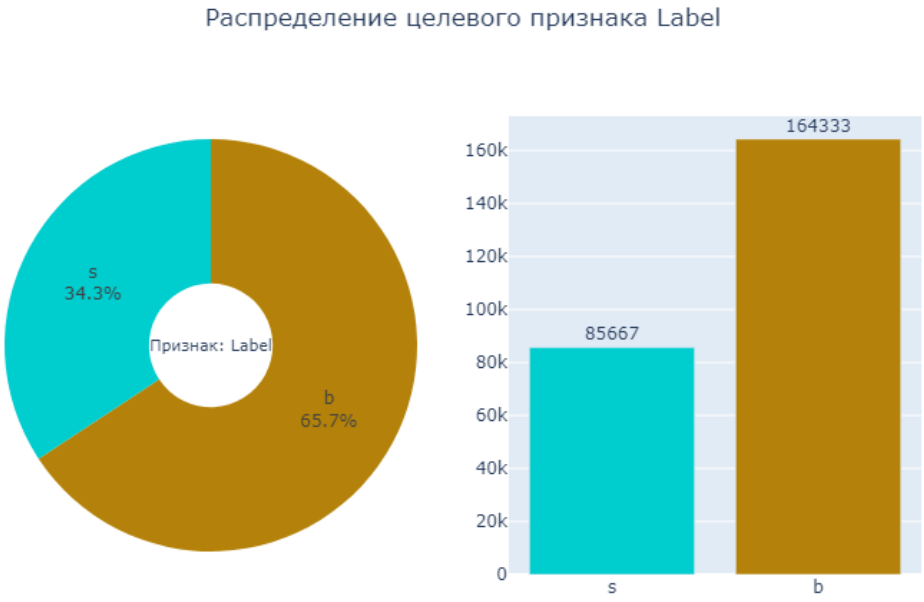
Name: PRI_jet_all_pt, count: 250000.000000, mean: 73.064591, std: 98.015662, min: 0.000000, 25%: -0.000000, 50%: 40.512500, 75%: 109.933750, max: 1633.433000,

Name: Weight, count: 250000.000000, mean: 1.646767, std: 1.875103, min: 0.001502, 25%: 0.018636, 50%: 1.156188, 75%: 2.404128, max: 7.822543;

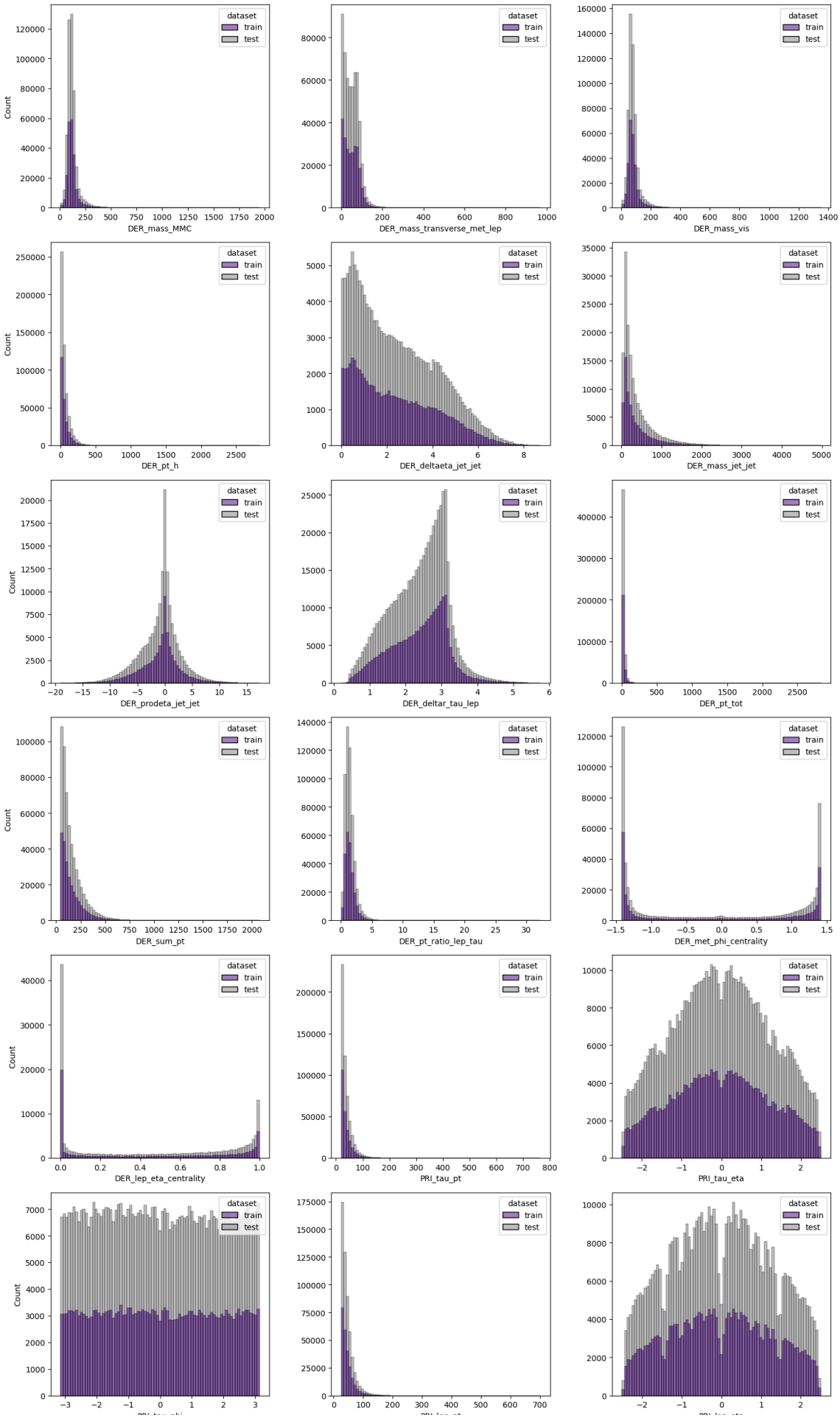
Name: Label, count: 250000, unique: 2, top: b, freq: 164333.

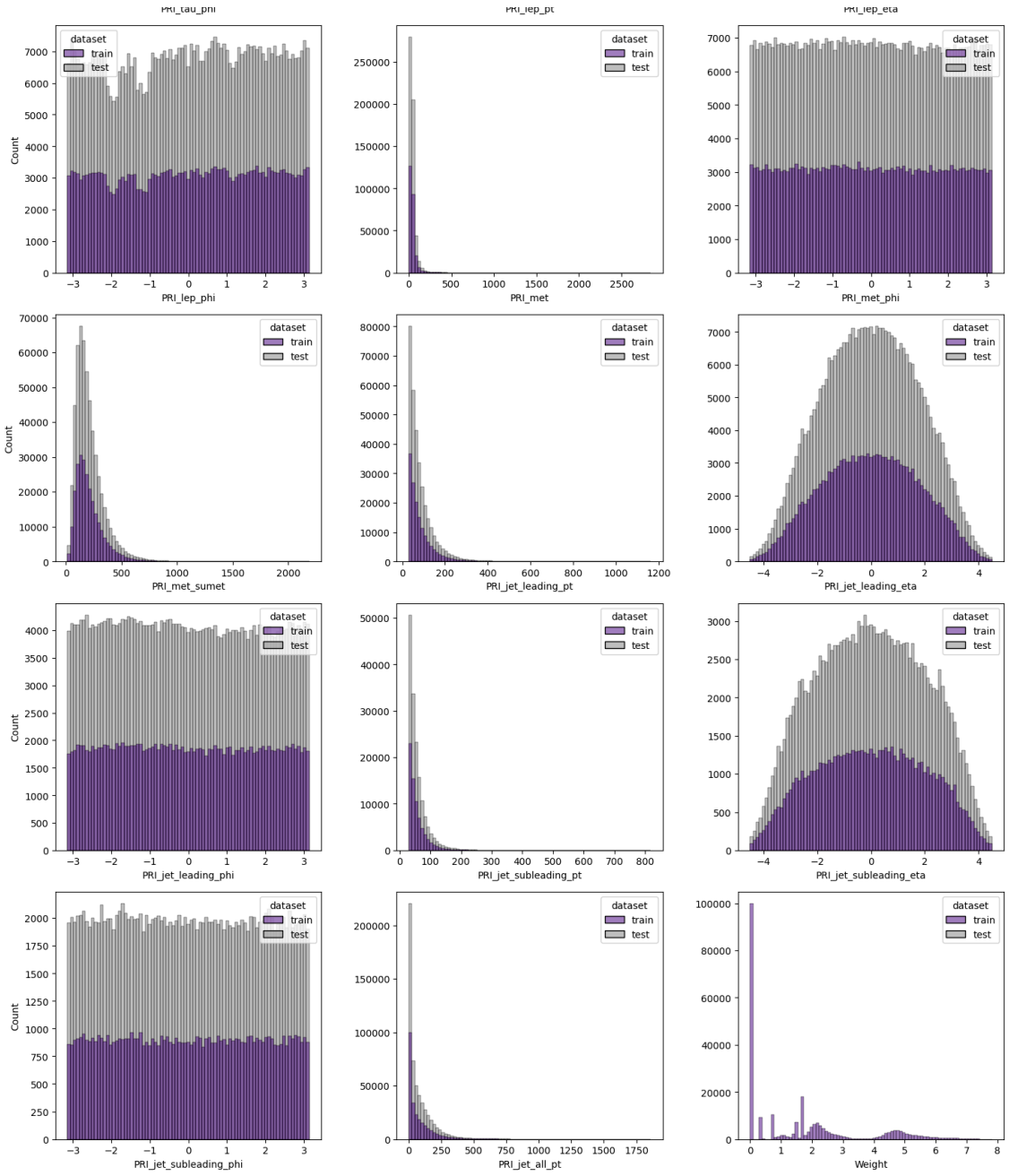
Анализ целевого признака

Целевой признак `Label` является бинарным и принимает значения `b` или `s`, указывающих на природу события (фоновое или сигнал).



Распределение остальных признаков (в тренировочном и тестовых наборах):


dataset
train
test
Count
DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality
PRI_tau_pt
PRI_tau_eta

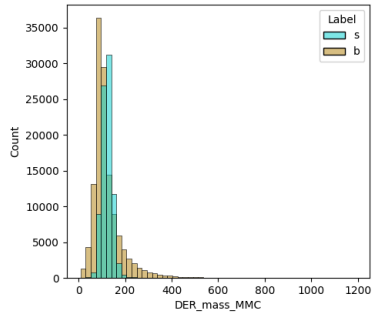


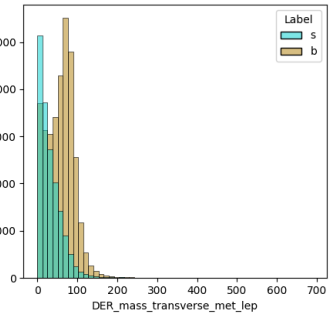
Сравнение распределений признаков по целевому классу в обучающей выборке

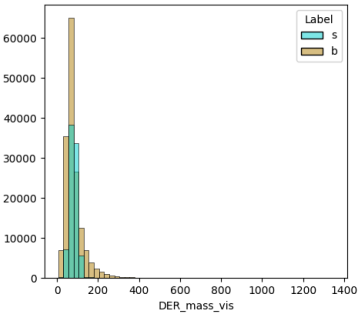
Далее мы сравниваем одномерные распределения признаков для фоновых событий и сигнальных событий в обучающем наборе.

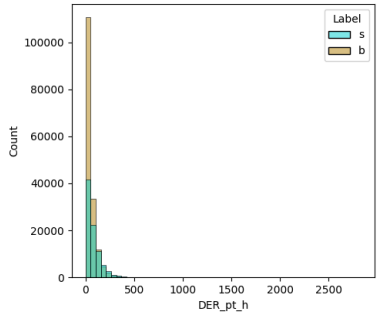
Если признак имеет разные распределения для фоновых и сигнальных событий, то это значит, что признак важен в задаче классификации событий, когда метка неизвестна.

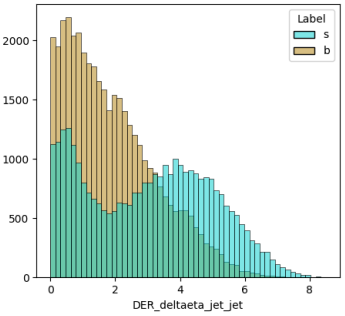
Аналогично, если признак имеет очень похожие распределения для двух целевых классов, то он вряд ли поможет в задаче классификации.

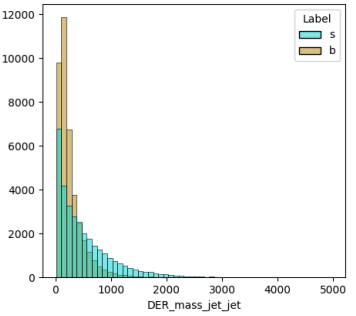

Label
s
b
Count
DER_mass_MMC

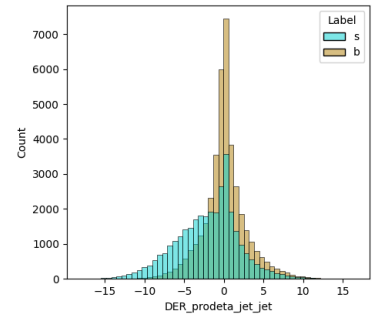

Label
s
b
DER_mass_transverse_met_lep

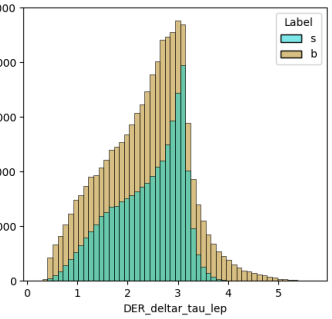

Label
s
b
DER_mass_vis

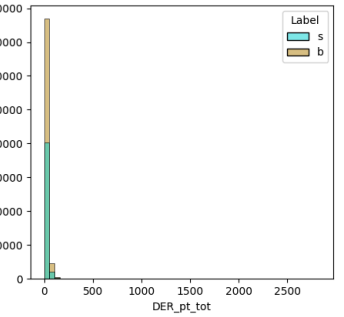

Label
s
b
Count
DER_pt_h

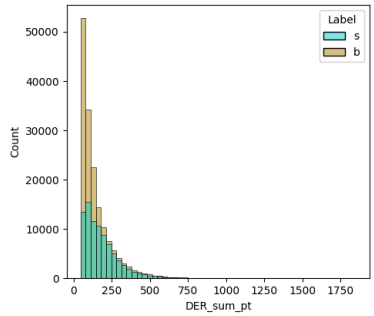

Label
s
b
DER_deltaeta_jet_jet

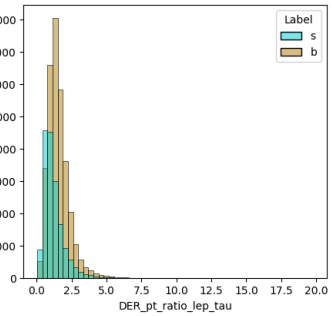

Label
s
b
DER_mass_jet_jet

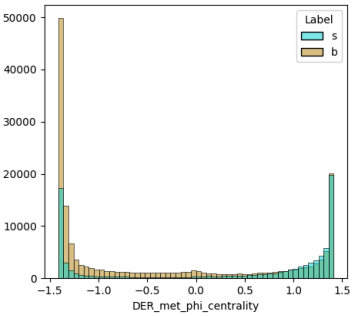

Label
s
b
Count
DER_prodeta_jet_jet

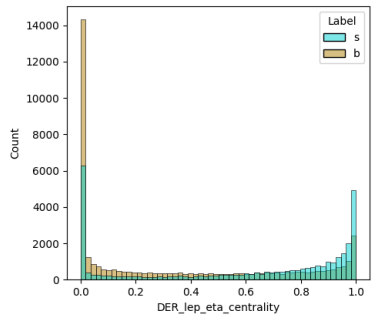

Label
s
b
DER_deltar_tau_lep

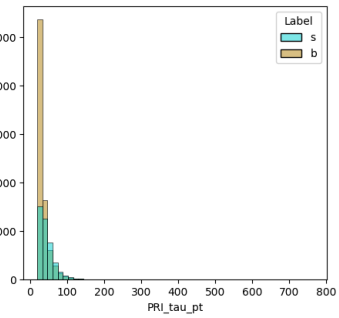

Label
s
b
DER_pt_tot

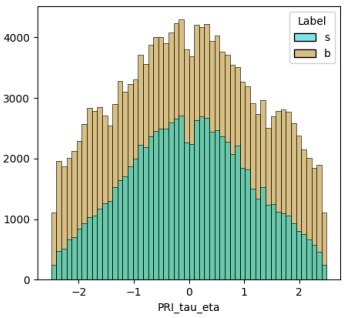

Label
s
b
Count
DER_sum_pt

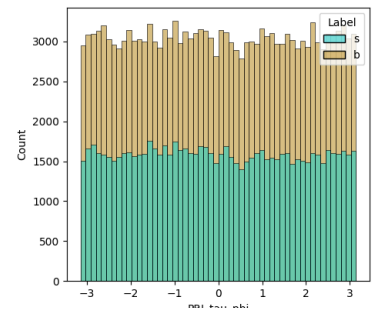

Label
s
b
DER_pt_ratio_lep_tau

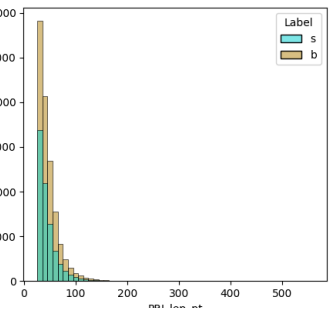

Label
s
b
DER_met_phi_centrality

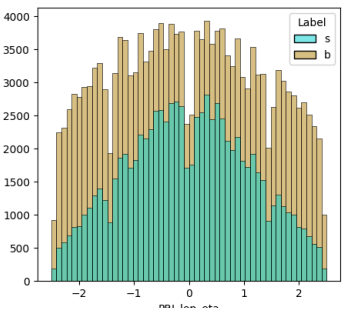

Label
s
b
Count
DER_lep_eta_centrality

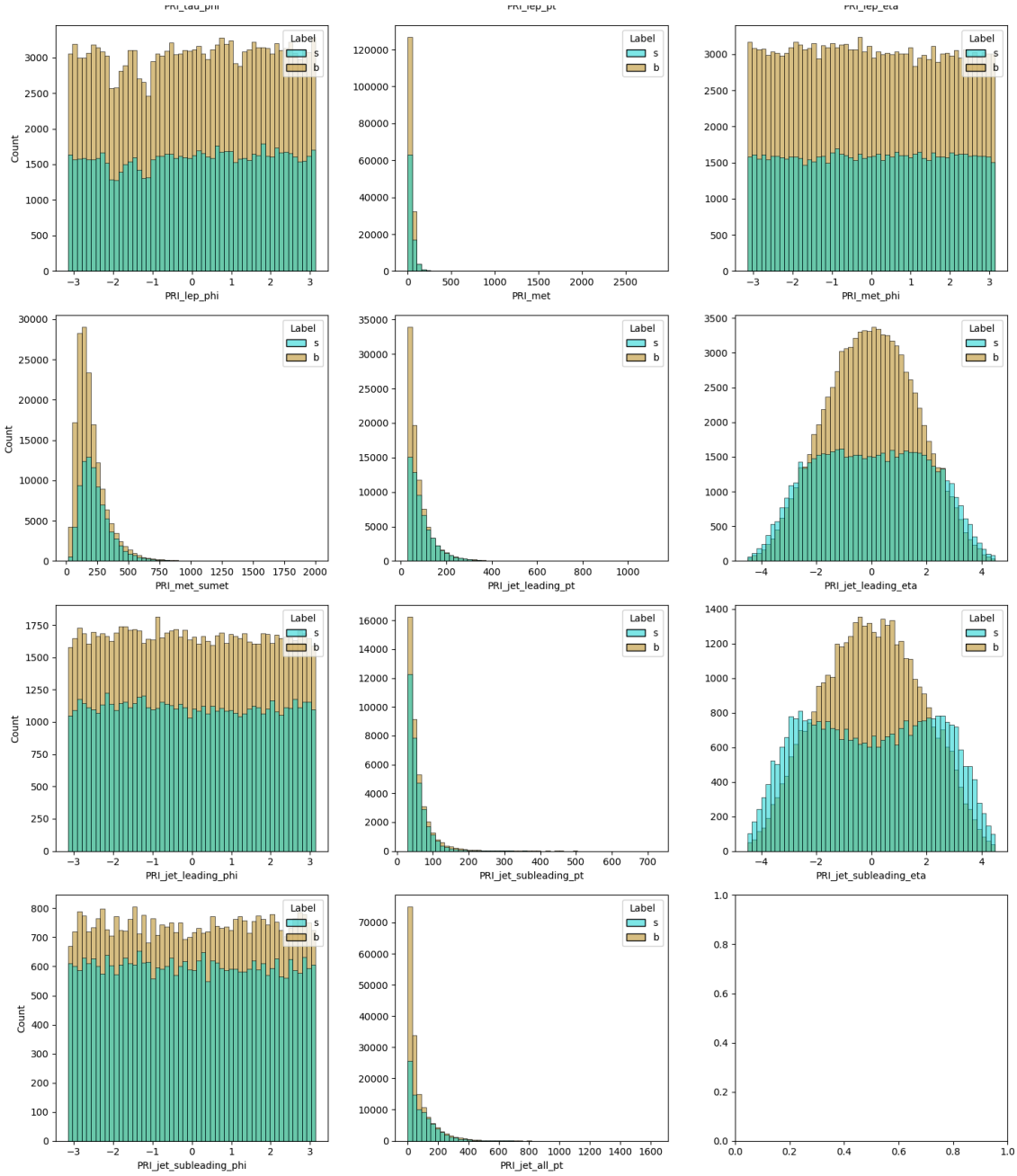

Label
s
b
PRI_tau_pt


Label
s
b
PRI_tau_eta

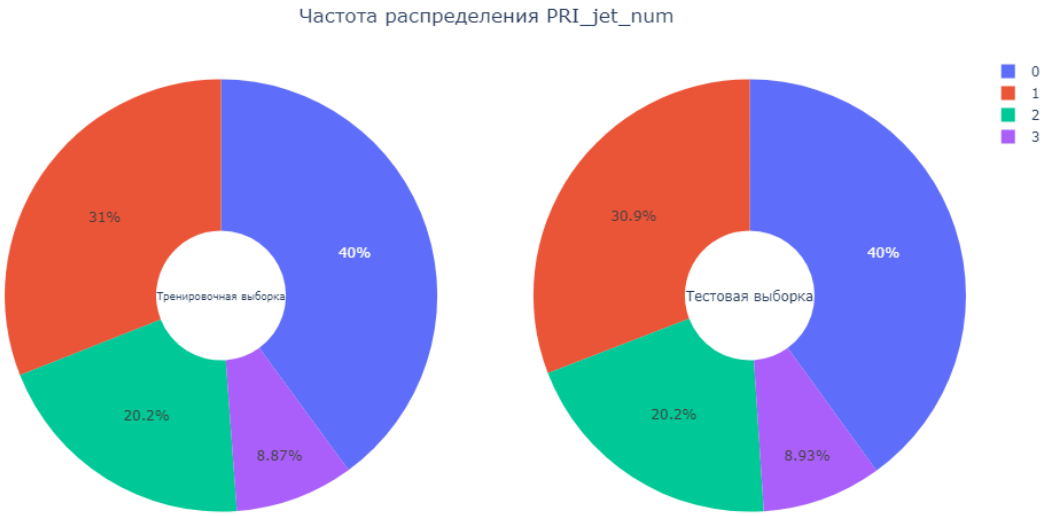

Label
s
b
Count

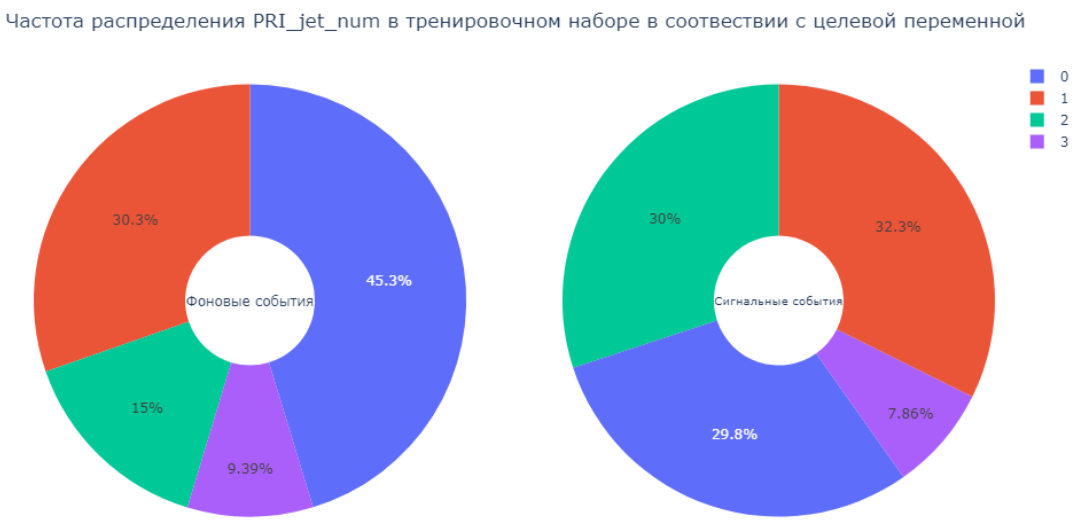

Label
s
b


Label
s
b



Распределение целочисленных признаков (количество струй PRI_jet_num):





Краткие выводы:

Были проанализированы стандартные параметры датасета (дубли, пустоты, типы данных признаков). Из описания данных были выявлены константные заполнения (-999) для пропущенных значений.

Был проведен анализ целевой переменной и построение ее распределения. Также были построены распределения признаков в разрезе целевой переменной и разрезе тренировочных и тестовых данных.

Судя по распределениям тренировочные и тестовые данных распределены одинаково, однако ряд признаков в тренировочном наборе имеет различные распределения в зависимости от целевой переменной (например DER_deltaeta_jet_jet, PRI_jet_subleading_eta, DER_lep_eta_centrality и тд). Это может свидетельствовать об особой значимости этих признаков для классификации.