Идентификация заряженных частиц в эксперименте LHCb

Евгений Елтышев

Кафедра Анализа данных
Факультет инноваций и высоких технологий
Московский физико-технический институт

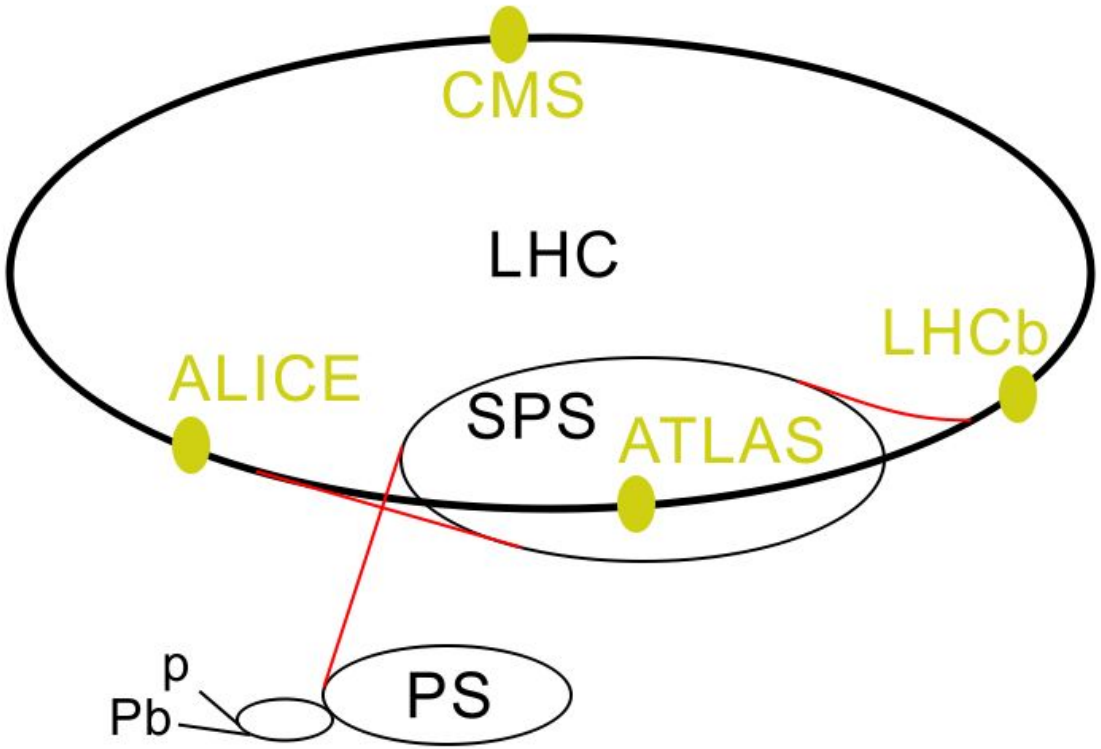
День X

План

1. Введение
   - Большой адронный коллайдер
   - Детектор LHCb
2. Описание задачи
   - Постановка задачи
   - Существующие решения
   - Данные и измерение качества
3. Результаты
   - Анализ особенностей данных
   - Новые модели
   - Будущие исследования

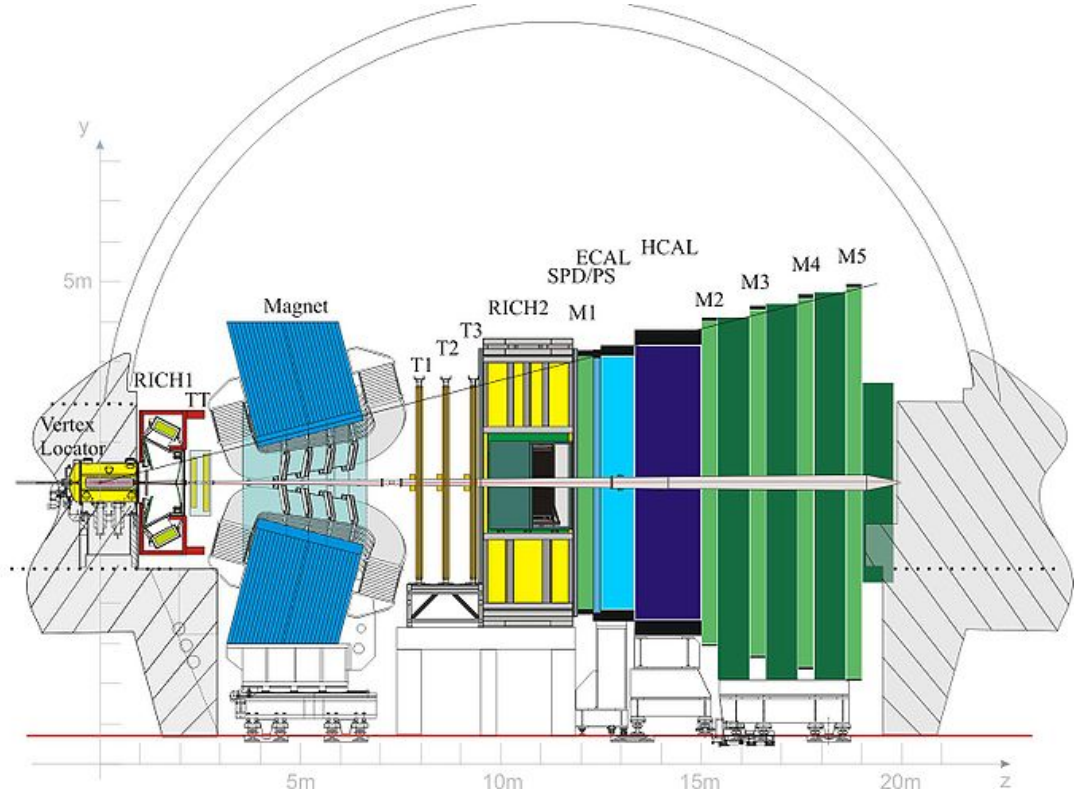
МФТИ

Большой адронный коллайдер



БАК - кольцо длиной 27км по которому циркулируют пучки протонов

Детектор LHCb



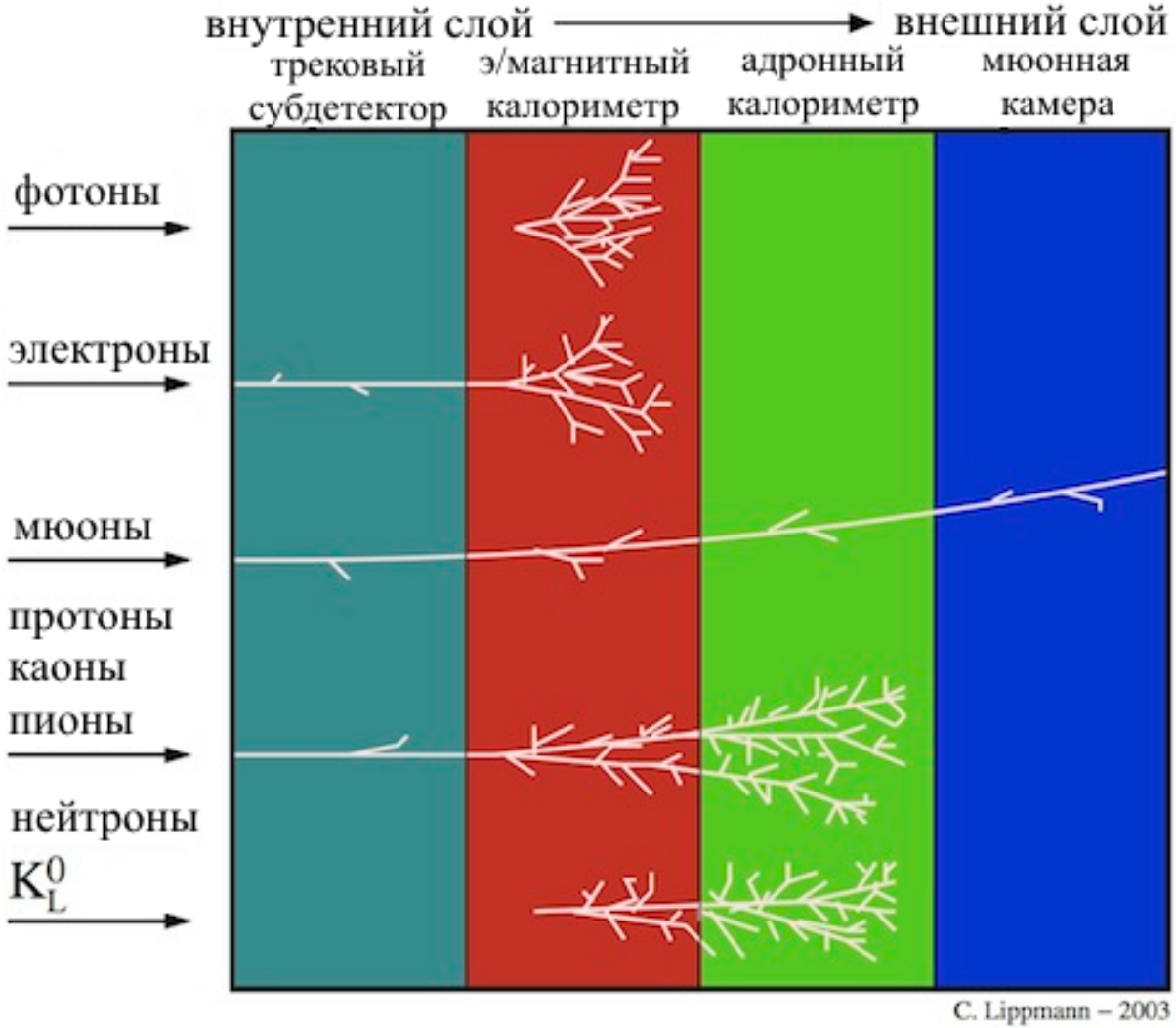
Последовательность слоев:

- Вершинный субдетектор (VELO)
- Черенковские счетчики (RICH-1, RICH-2)
- Трековый субдетектор (TT, T1-T3)
- Калориметры (ECAL, HCAL)
- Мюонная камера (M1-M5)

Постановка задачи

По характеристикам трека определить вид частицы:

- Ghost
- Электрон
- Мюон
- Пион
- Каон
- Протон



Существующие решения

Delta log-likelihood models (DLL)

Строится 6 вероятностных моделей на основе характеристик каждой из частиц с точки зрения физики. Выбор вида частицы производится по максимуму правдоподобия

ProbNN

- Специализированная библиотека для анализа данных в физике TMVA
- Перцептрон с одним скрытым слоем, 6 моделей "один против всех"
- Включает в себя как DLL-признаки, так и низкоуровневую информацию

МФТИ

Данные

77 признаков

- **Характеристики трека:** импульс, заряд и т.д.
- **Правдоподобия от различных слоев:** CaloDLL, RichDLL, MuonLL, CombDLL
- **Геометрическая информация:** положение исходной вершины
- **Флаги прохождения частицы через слои:** через RICH, CALO, Muon

Выборки

- train – 1.2 млн треков (стандартные пропорции классов)
- test – 1 млн треков (стандартные пропорции классов)
- equal mix – 300 тыс. треков (сбалансированные классы)

МФТИ

Измерение качества

Базовое решение

Базовое решение имеет вид 6 моделей "один против всех". Его качество измерялось как 6 значений ROC AUC.

Мои исследования

Рассматривались не только "один против всех"модели, но и многоклассовые.

- ROC AUC "один против всех"
- Log Loss - как целевая функция для многоклассовых моделей

МФТИ

Проделанная работа

Проанализированы и устранены особенности данных

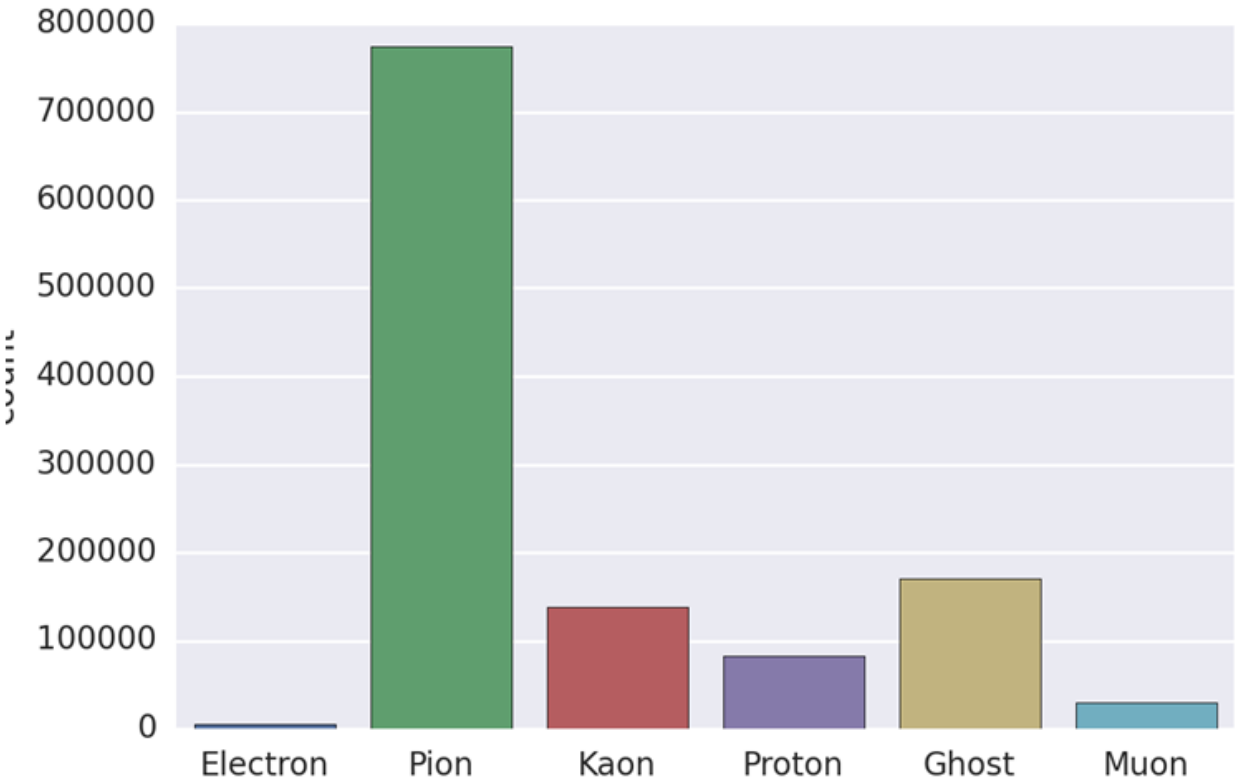
- Несбалансированность классов
- Длинные хвосты распределений
- Пропуски в данных

Использованы новые модели

- Логистическая регрессия
- XGBoost
- Нейросети Keras
- Специальная архитектура нейросети

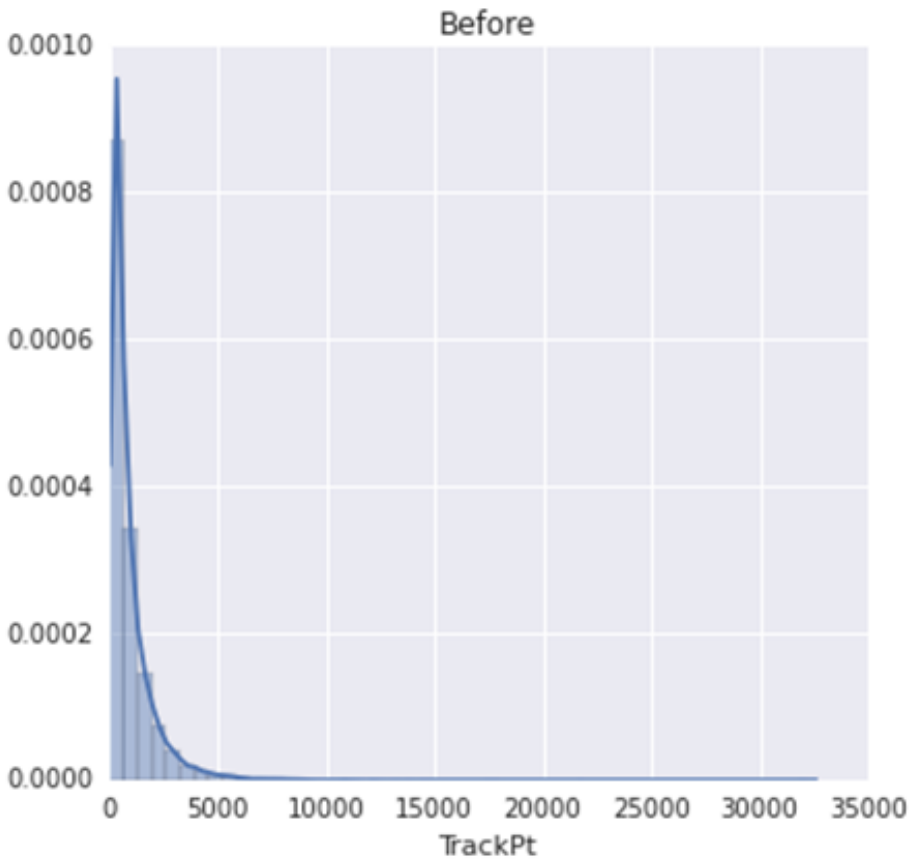
МФТИ

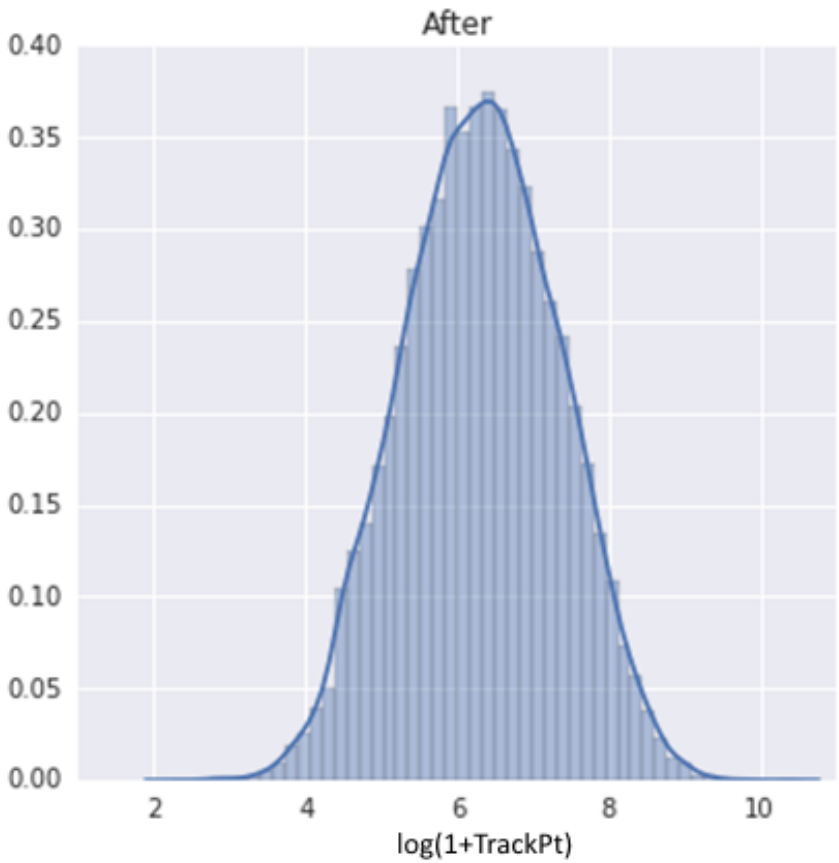
Несбалансированность классов



- Классы крайне несбалансированы
- Градиентные методы плохо настраиваются на маленькие классы
- Как исправить: взвешивание выборки, downsampling

Длинные хвосты распределений





Длинные хвосты распределений плохо влияют на линейные методы

Для "сглаживания"распределений можно применить $sign(x)\log(1+|x|)$

МФТИ

Пропуски в данных

- В данных много пропусков, причем они заменены на -999
- Не мешает методам, основаным на деревьях
- Мешает линейным методам
- Для линейных методов: заменим пропуски на среднее значение в колонке, добавим бинарный признак "был ли пропуск"

Логистическая регрессия

Обычная логистическая регрессия из пакета scikit-learn с L1-регуляризацией

Частица	Baseline	Логистическая регрессия
Ghost	**5.2**	5.64
Электрон	**1.73**	1.88
Мюон	1.21	**1.14**
Пион	6.65	**6.56**
Каон	**8.48**	9.57
Протон	**8.39**	10.23

Ошибка: (1-AUC)*100

XGBoost

- Мощная библиотека, реализующая градиентный бустинг над деревьями
- Многоклассовая классификация вместо стратегии "один против всех"

Частица	Baseline	XGBoost
Ghost	5.2	**4.4**
Электрон	1.73	**1.2**
Мюон	1.21	**0.6**
Пион	6.65	**4.8**
Каон	8.48	**7.1**
Протон	8.39	**7.4**

Ошибка: (1-AUC)*100

Нейросети Keras

- Библиотека для нейросетей, основанная на Theano
- Однослойный перцептрон, скрытых узлов - 1.5*количество фичей

Частица	Baseline	Keras
Ghost	5.2	N/A
Электрон	1.73	N/A
Мюон	1.21	N/A
Пион	6.65	N/A
Каон	8.48	N/A
Протон	8.39	N/A

Ошибка: (1-AUC)*100

МФТИ

Специальная структура нейросети

Картинка с архитектурой

Специальная структура нейросети

Частица	Baseline	BlockNN
Ghost	5.2	N/A
Электрон	1.73	N/A
Мюон	1.21	N/A
Пион	6.65	N/A
Каон	8.48	N/A
Протон	8.39	N/A

Ошибка: (1-AUC)*100

Сравнение моделей

Частица	Baseline	Логистическая регрессия	XGBoost	KerasNN	BlockNN
Ghost	5.2	5.64	**4.4**	N/A	N/A
Электрон	1.73	1.88	**1.2**	N/A	N/A
Мюон	1.21	1.14	**0.6**	N/A	N/A
Пион	6.65	6.56	**4.8**	N/A	N/A
Каон	8.48	9.57	**7.1**	N/A	N/A
Протон	8.39	10.23	**7.4**	N/A	N/A

Ошибка: (1-AUC)*100

Будущие исследования

Равномерность модели

Физическая постановка задачи накладывает ограничение равномерности на модель: качество модели должно быть примерно одинаковое при всех значениях одной из переменных - импульса - и при низких, и при высоких.

Онлайн\Оффлайн предсказания

Идентификация частиц используется, грубо говоря, в двух режимах:

- Онлайн - для отсева событий "на лету". Ограничения - 10-20 мс на предсказание
- Оффлайн - при постобработке результатов, скорость не так сильно важна. Важно качество

МФТИ

Результаты, выносимые на защиту

Произведена работа с данными

- Сбалансированы классы
- Сглажены распределения некоторых признаков
- Устранены пропуски в данных

Повышено качество предсказаний

- XGBoost - ошибка снизилась в среднем на 25%
- Нейросети Keras - ошибка снизилась в среднем на X %

Хороший фундамент для будущих исследований

- Модель XGBoost может использоваться в оффлайн предсказаниях для достижения высокого качества
- Нейронные сети могут использоваться в онлайн предсказаниях из-за крайне высокого быстродействия