



Применение методов машинного обучения для идентификации струй, образованных W -бозоном

Научный руководитель:

к.ф.-м.н. Солдатов Евгений Юрьевич

Научный консультант:

к.ф.-м.н. Мягков Алексей Григорьевич

Студент:

Ван Алина Маошэновна, М23-112

Москва, 2025

Целью работы является применение методов машинного обучения для решения задачи идентификации струй, образованных W -бозоном, в процессе образования пары топ-кварков.

Задачи:

1. Отбор событий для последующего формирования обучающих и тестовых выборок из Монте-Карло данных с учетом веса каждого события;
2. Анализ оптимальных гиперпараметров и обучение моделей MLP и BDT;
3. Тестирование и оценка моделей на Монте-Карло данных;
4. Исследование отклика обученных моделей на реальных данных;
5. Сравнение эффективности идентификации струй от W -бозона при помощи обученных моделей и при помощи переменной `wtag_mass_d2`.

Исходные данные. Отборы

Датасет с сигналом строится на основе Монте-Карло симуляции процесса образования пары топ кварков.

Отборы

- Как минимум одна b-меченная струя
- Условие на кинематику струи
 - $p_t > 30 \text{ GeV}$
 - $\eta < 2.5$

Датасет с фоном строится на основе Монте-Карло симуляций процессов:

- $V V + \text{КХД-струи}$
- $W(\rightarrow e\nu) + \text{КХД-струи}$
- $Z(\rightarrow ee) + \text{КХД-струи}$
- $Z(\rightarrow \tau\tau) + \text{КХД-струи}$
- $W(\rightarrow \nu\tau) + \text{КХД-струи}$
- $\text{Single-top} + \text{КХД-струи}$

На **фоновые и сигнальные** события наложены ограничения на условие отбора для толстой струи:

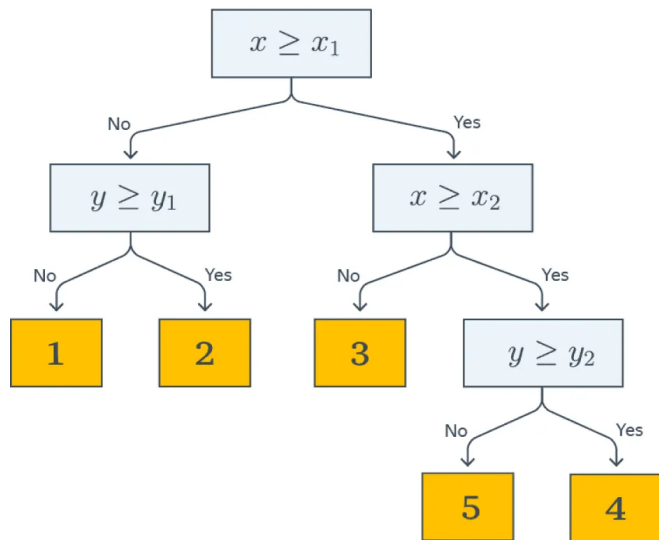
$$60 \text{ GeV} < M < 110 \text{ GeV}$$

$$p_t > 200 \text{ GeV}$$

Используются данные RUN2 2017 года с **эксперимента ATLAS** со светимостью 6.1 fb^{-1} и энергией 13 ТэВ.

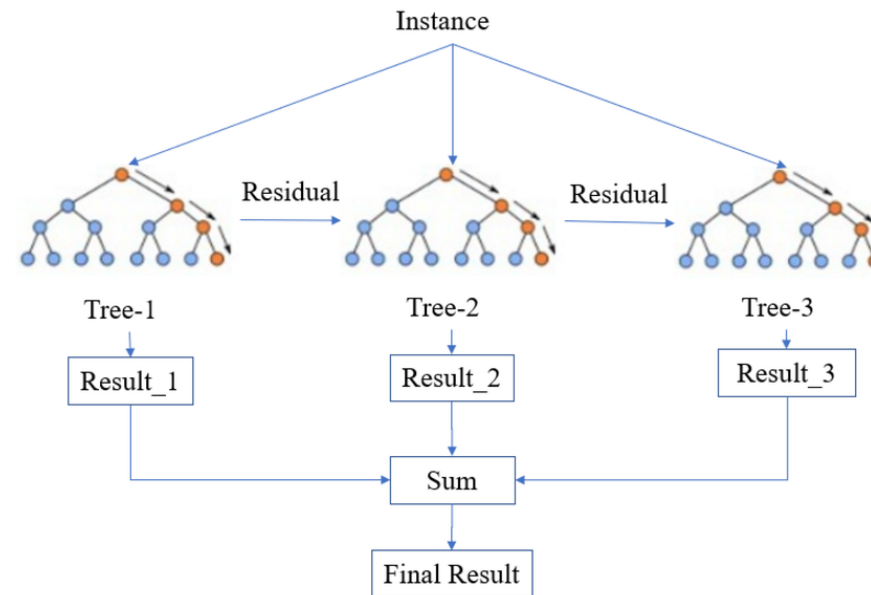
Бустированные деревья решений (BDT)

Дерево решений



Разбивает пространство признаков на области по особому *правилу* и приписывает этим областям *метку класса*.

Градиентный бустинг

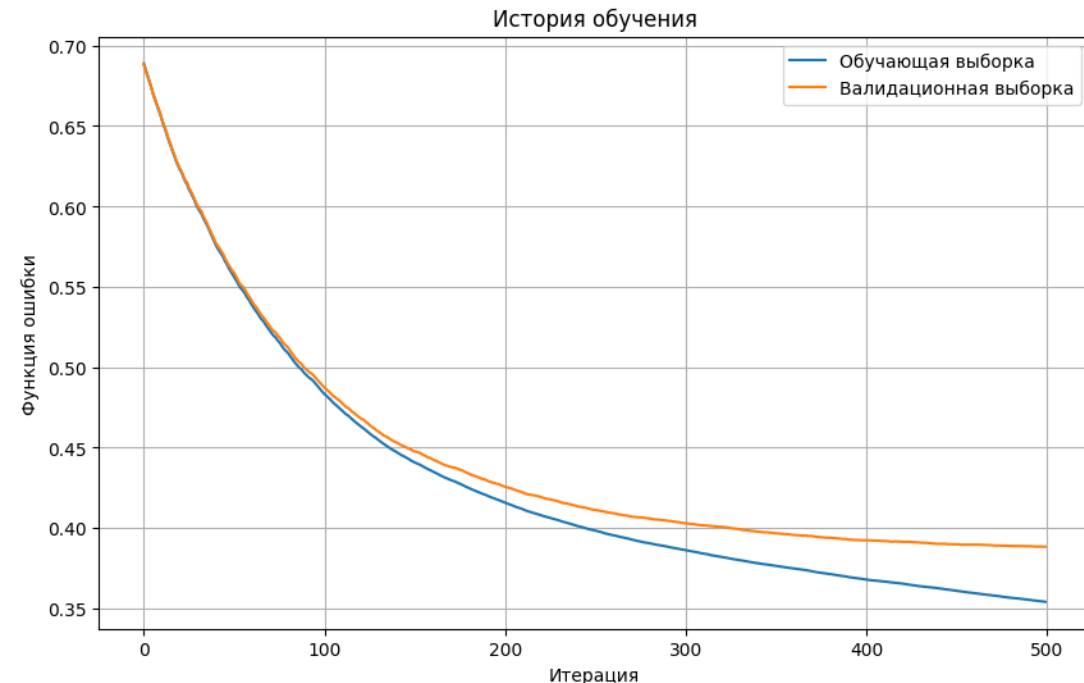


Последовательно обучает *слабые* модели на ошибках *предыдущих* моделей.

Обучение и настройка гиперпараметров модели BDT

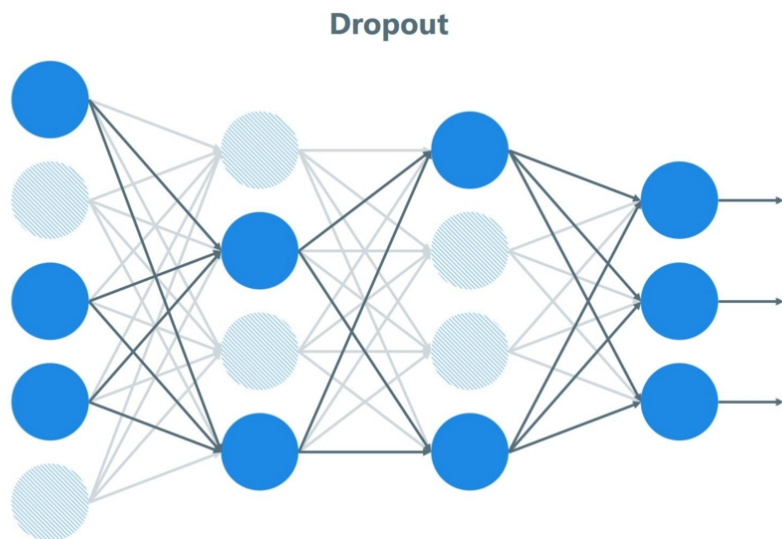
Для настройки гиперпараметров использовался жадный алгоритм **поиска по сетке**.

Гиперпараметр	Значение
Максимальная глубина	3
Минимальный вес для дочерних узлов	3
Скорость обучения	0.01
Доля выборки для обучения дерева	0.6
Доля признаков для построения дерева	0.6
Параметр L2-регуляризации	0.5
Параметр L1-регуляризации	0.1
Функция потерь	logloss



Обучение производилось с данными по всем, перечисленным ранее, фоновым процессам с учетом *веса каждого события*

Многослойный персептрон (MLP)



Методы решения проблемы переобучения:

- L1, L2 - регуляризация
- Dropout

Нейронная сеть — это дифференцируемая сложная функция, преобразующая пространство признаков в пространство откликов.

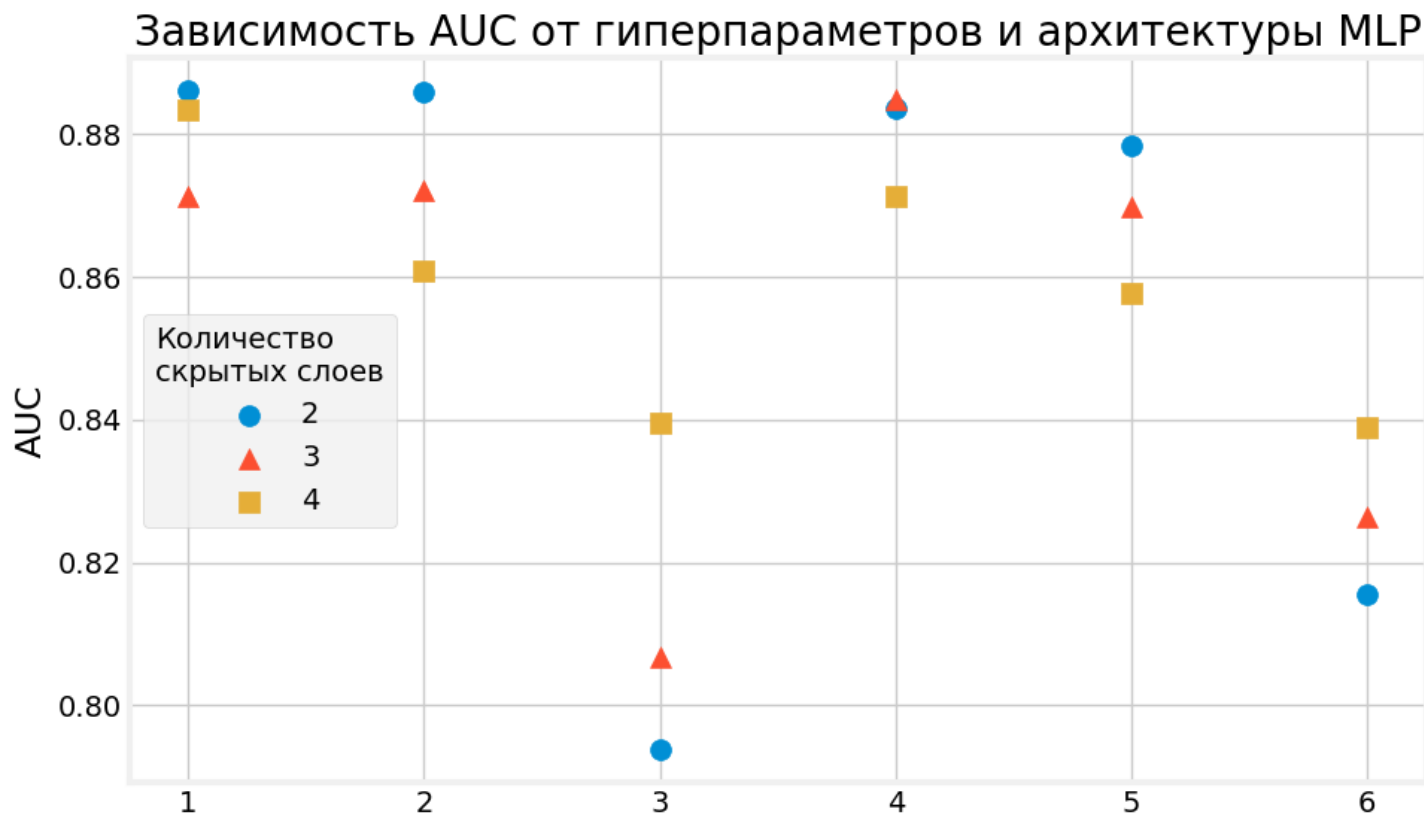
Выходные данные полносвязного слоя формально можно представить:

$$y = \sigma(Wx + b)$$

Ключевые механизмы обучения:

1. Прямой проход
2. Вычисление взвешенной* функции потерь
3. Обратное распространение ошибки

Настройка гиперпараметров модели MLP



Вариации гиперпараметров

№	Батч	Скорость обучения
1	128	10^{-3}
2	128	10^{-4}
3	128	10^{-5}
4	64	10^{-3}
5	64	10^{-4}
6	64	10^{-5}

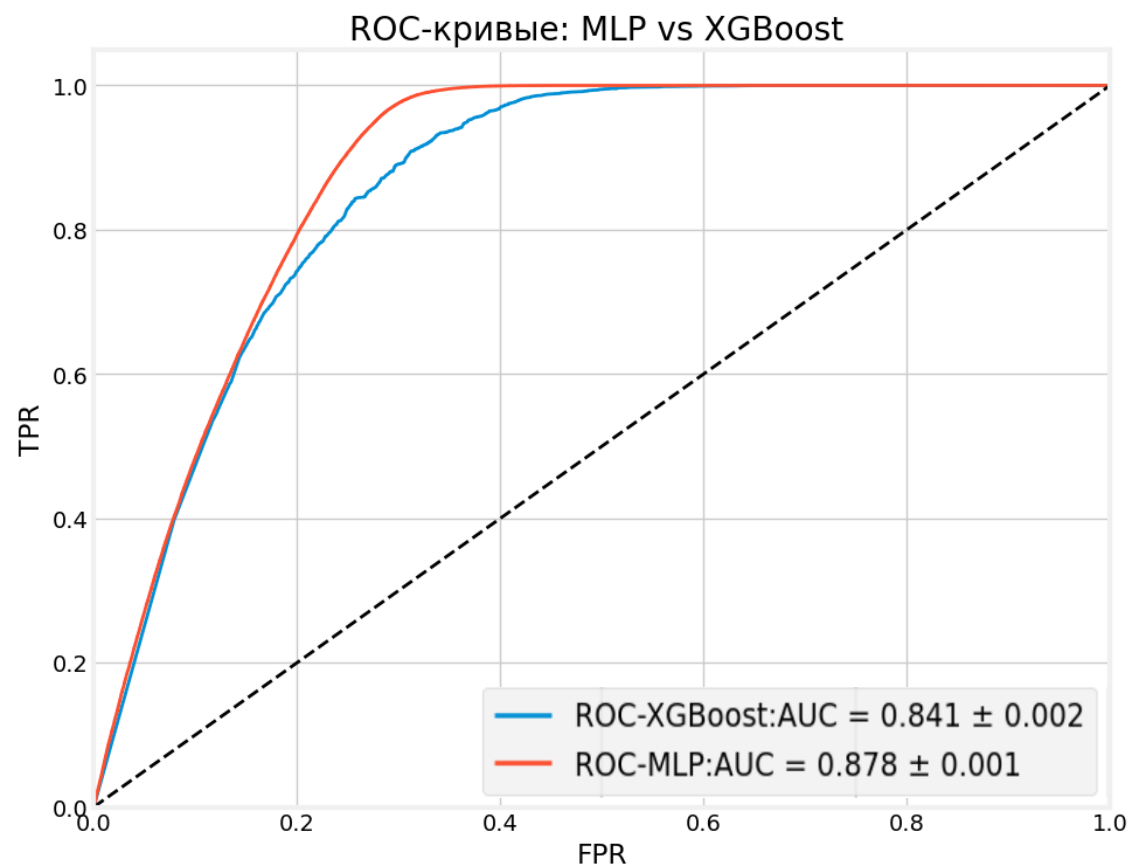
Итоговая архитектура MLP



Слой	Характеристика
Входной	13
Скрытый	32
Dropout	0.1
Скрытый	64
Dropout	0.1
Выходной	1

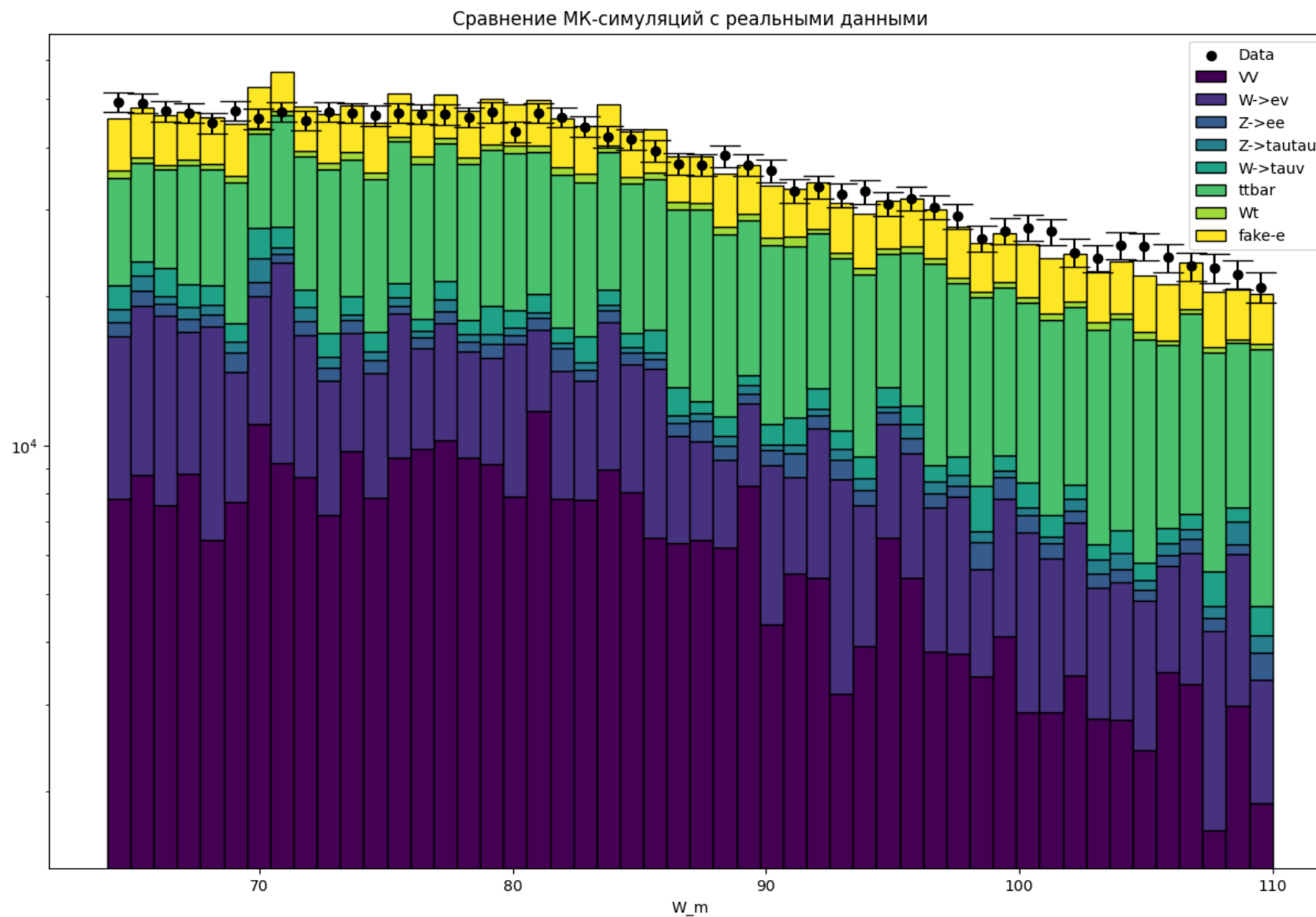
Параметр	Значение
Скорость обучения	0.001
Батч	128
Функция активации	ReLU
Оптимизатор	RMSPProp
Кол-во эпох	20

Тестирование моделей MLP и BDT на Монте-Карло данных



Модель	Accuracy	AUC
MLP	0.860 ± 0.001	0.878 ± 0.001
BDT	0.836 ± 0.001	0.841 ± 0.002

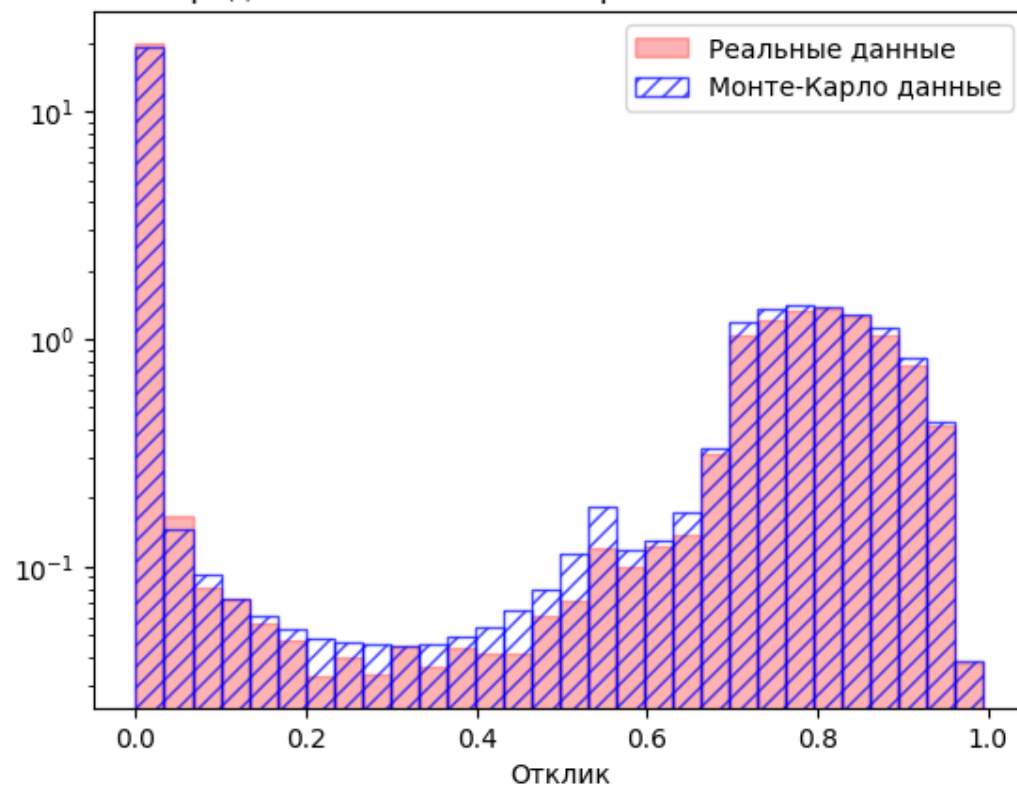
Сравнение Монте-Карло процессов с реальными данными



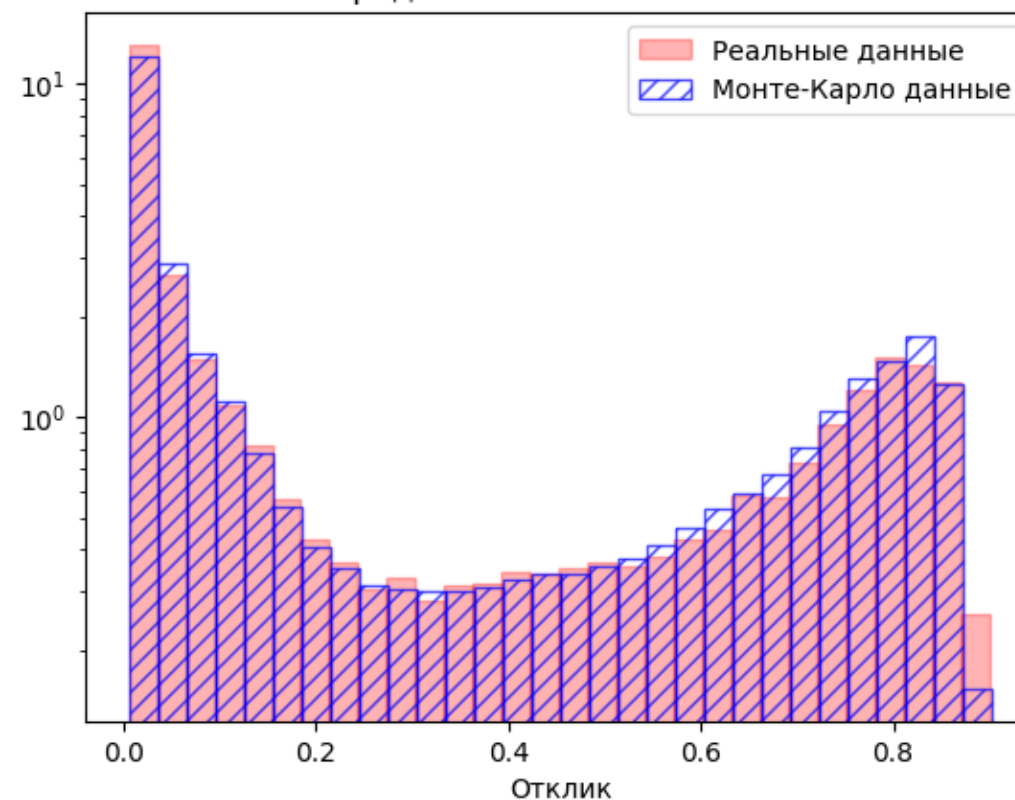
Исследование отклика моделей на реальных данных



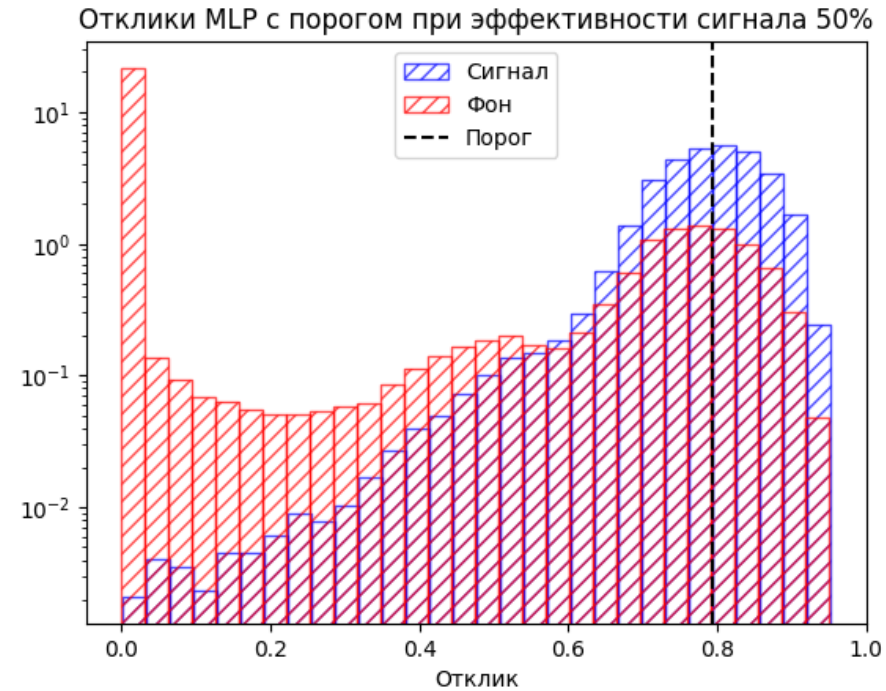
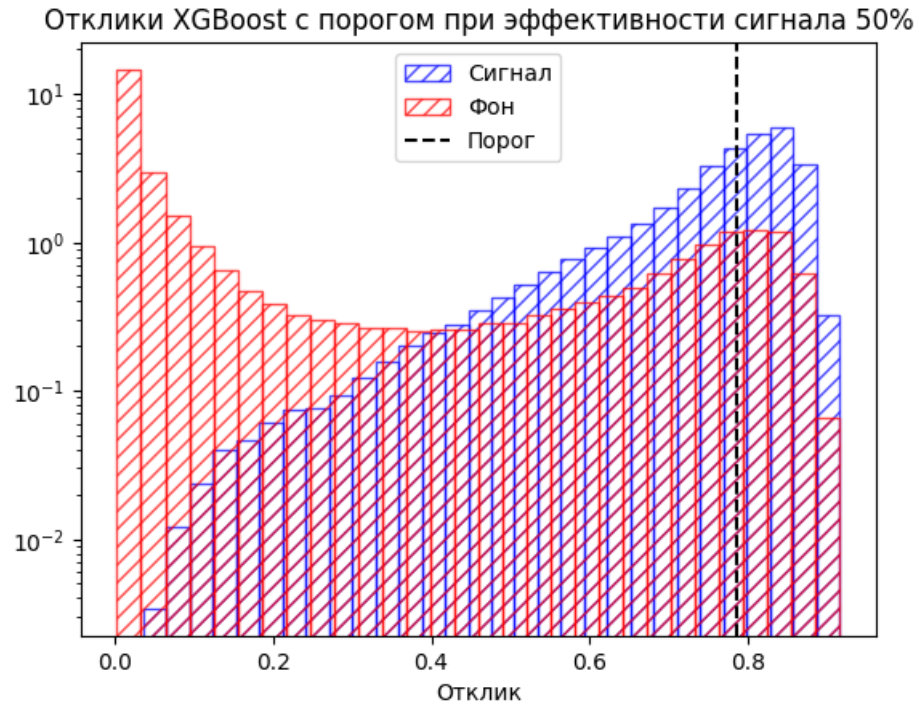
Распределение плотности вероятности откликов MLP



Распределение откликов XGBoost



Сравнение эффективности идентификации W-бозона между моделями и wtag_mass_d2



Классификатор	Эффективность фона
MLP	10.5 % $\pm$ 0.1 %
BDT	10.6 % $\pm$ 0.1 %
wtag_mass_d2	30.9 % $\pm$ 0.1 %

Заключение

В рамках работы были сделаны **следующие задачи**:

1. Сформированы датасеты с входными переменными из МК-процессов и реальных данных.
2. Проведены настройка архитектуры и обучение моделей машинного обучения MLP и BDT.
3. Полученные точности моделей на тестовой выборке:

Модель	Ассурасу
MLP	0.860 ± 0.001
BDT	0.836 ± 0.001

4. Проведен анализ откликов моделей на реальных данных.
5. При заданной эффективности сигнала 50% посчитаны эффективности фона для двух классификаторов и дискриминирующей переменной.

Классификатор	Эффективность фона
MLP	$10.5 \% \pm 0.1 \%$
BDT	$10.6 \% \pm 0.1 \%$
wtag_mass_d2	$30.9 \% \pm 0.1 \%$

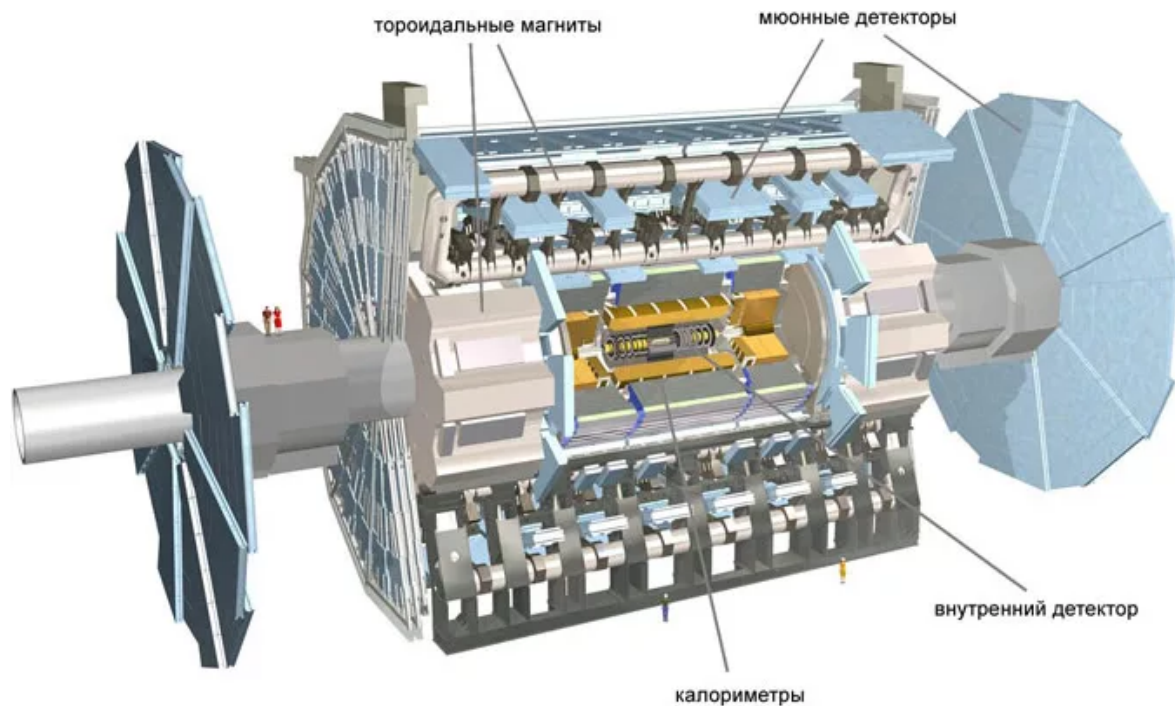
Спасибо за внимание!



Дополнительные слайды



Детектор ATLAS



Устройство детектора:

- Внутренний детектор:
 - Пиксельный полупроводниковый детектор - точная реконструкция вершин взаимодействия
 - Кремниевый микроstriповый детектор
 - Детектор переходного излучения
- Соленоидальный магнит:
- Электромагнитный калориметр
- Адронный калориметр
- Мюонная система

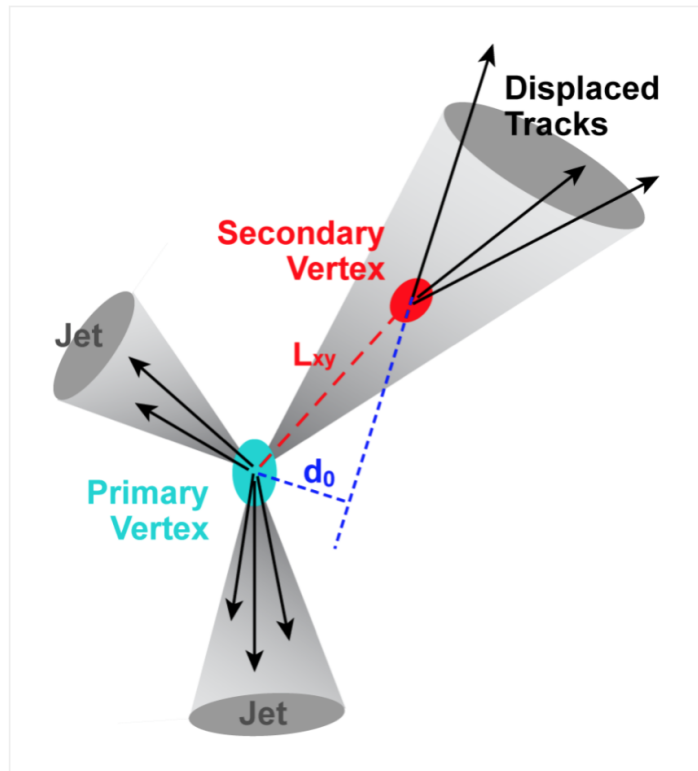
Триггеры:

- Низкоуровневый триггер L1
- Высокоуровневый триггер HLT

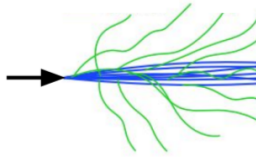
b-меченные струи. Толстые струи

b-мечение основано на свойствах b-адронов:

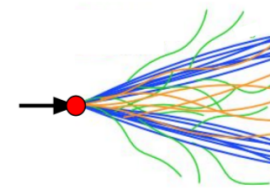
- Смещенная вершина (малое время жизни)
- Большая масса



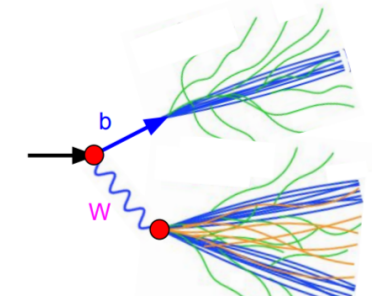
Quark/Gluon



Higgs/W/Z Bosons

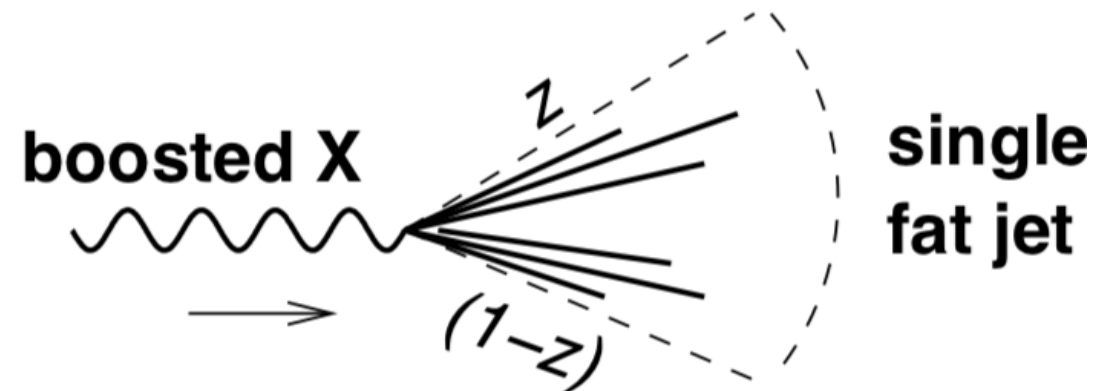


Top Quark



Толстые струи

При высоких p_t бозона струи имеют малый угол разлета. Струи попадают в конус $R = 1$.



Нормализация

Преобразование **Йео-Джонсона** служит для стабилизации дисперсии и нормализации распределения данных как с положительными, так и с отрицательными значениями.

$$\psi(\lambda, y) = \begin{cases} ((y+1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Преобразование параметризуется с помощью **λ -параметра**, которое варьируется для наилучшего приближения к нормальному распределению.

Веса. Количество событий в выборках

$$W = \frac{L \times \sigma \times k_{fac} \times \epsilon_f}{sum_of_weight \times prw_weight}$$

Обучающая выборка	1600000
Валидационная выборка	480000
Тестовая выборка	480000

Входные переменные (1)

- Коэффициенты энергетической корреляции

$$E_{CF1} = \sum_i^n p_{T,i}, \quad E_{CF2} = \sum_{i,j}^n p_{T,i} p_{T,j} \Delta R_{ij}, \quad E_{CF3} = \sum_{i,j,k}^n p_{T,i} p_{T,j} p_{T,k} \Delta R_{ij} \Delta R_{jk} \Delta R_{ki}$$

$$D_2 = E_{CF3} \times \left(\frac{E_{CF1}}{E_{CF2}} \right)^3, \quad C_2 = \frac{E_{CF1} \times E_{CF2}}{E_{CF3}^2}$$

- Плоский поток

$$I_{\omega}^{kl} = \frac{1}{m_{jet}} \sum_i \omega_i \frac{p_{i,k}}{\omega_i} \frac{p_{i,l}}{\omega_i},$$

$$P = \frac{4\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2}$$

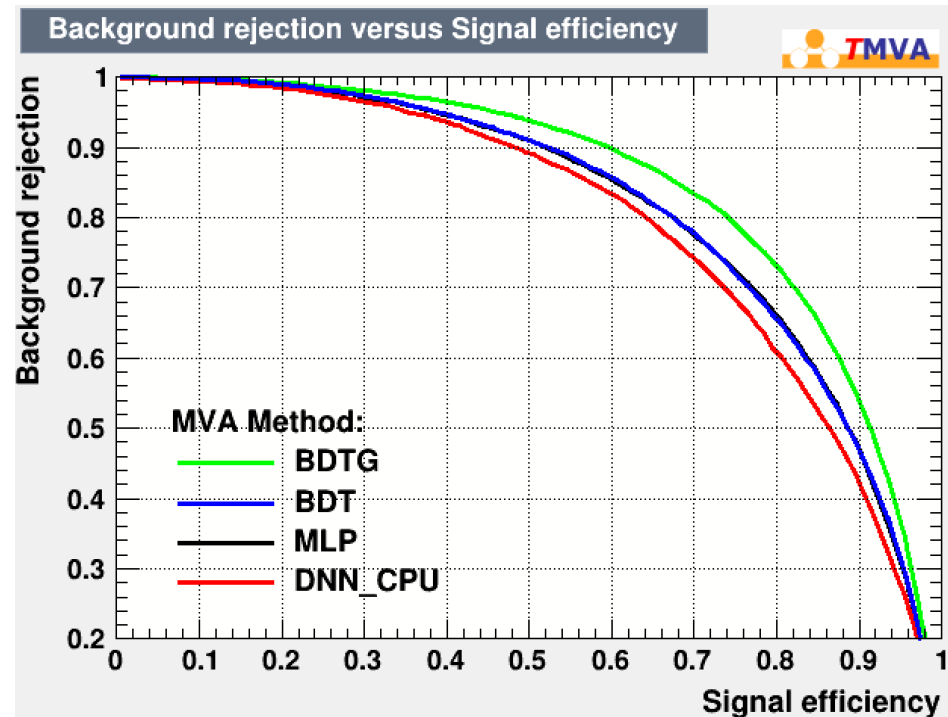
- Апланарность

$$S^{\alpha,\beta} = \frac{\sum_i p_i^{\alpha} p_i^{\beta}}{\sum_i |\vec{p}_i|^2}, \quad A = \frac{3\lambda_3}{2}$$

Входные переменные (2)

- Момент Фокса-Вольфрама
$$H_l = \sum_i \frac{|\vec{p}_i| |\vec{p}_j|}{E^2} P_l(\cos \theta_{ij})$$
- N-субъектность
$$\tau_N^{(\beta)} = \sum_i p_{T,i} \min(R_{1,i}, R_{2,i}, \dots, R_{N,i})$$
- Статистическая мера KtDR
$$\omega_{ij} = \exp \left\{ -\alpha \frac{d_{ij} - d^{min}}{d^{min}} \right\},$$
$$KtDR = \frac{RMS}{\langle m \rangle}.$$
- Мера разделения Z_{cut}
$$z_{cut} \equiv \frac{d_{cut}}{d_{cut} + Q_M^2}$$

Изученные методы МО



Используемые модели:

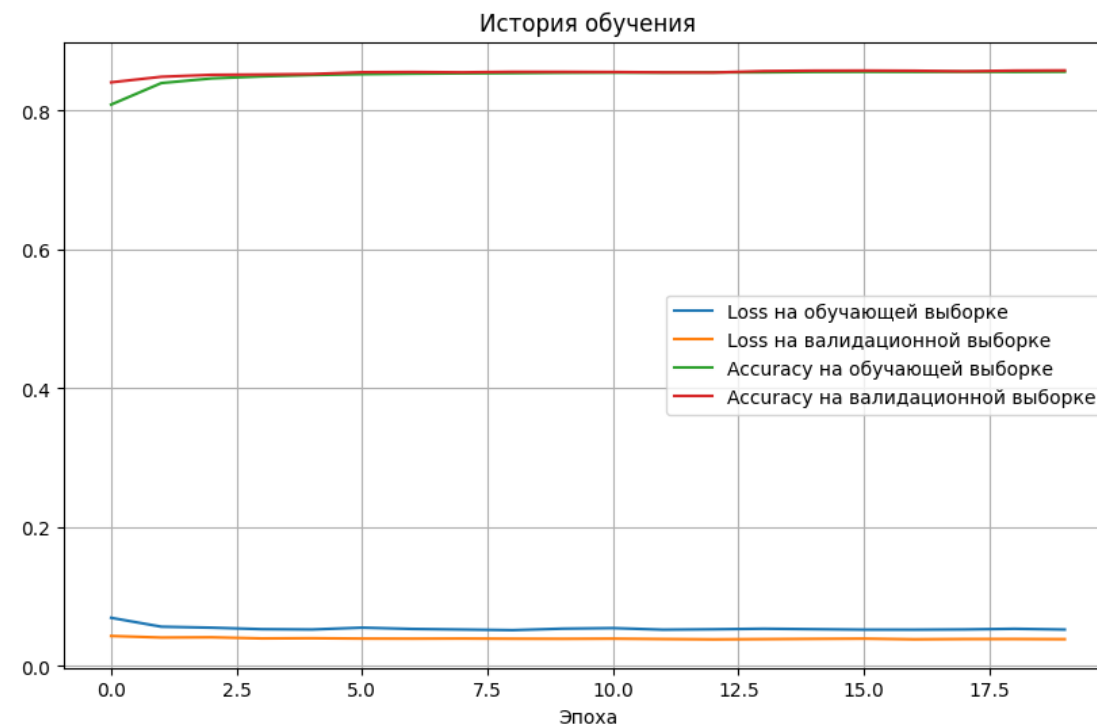
- **Ансамблевые модели. Бустинг:**
 - AdaBoost
 - Градиентный бустинг
- **Нейронные сети:**
 - Сверточные нейронные сети
 - Персептрон

Методы регуляризации:

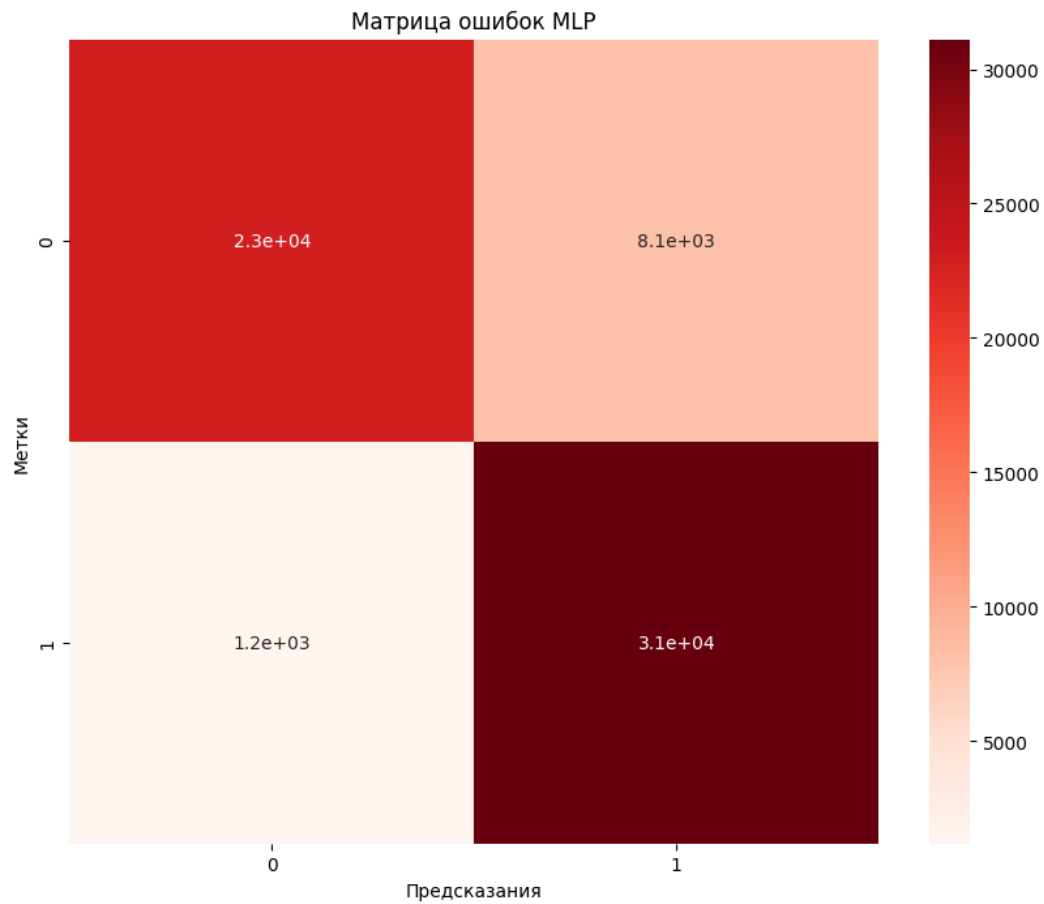
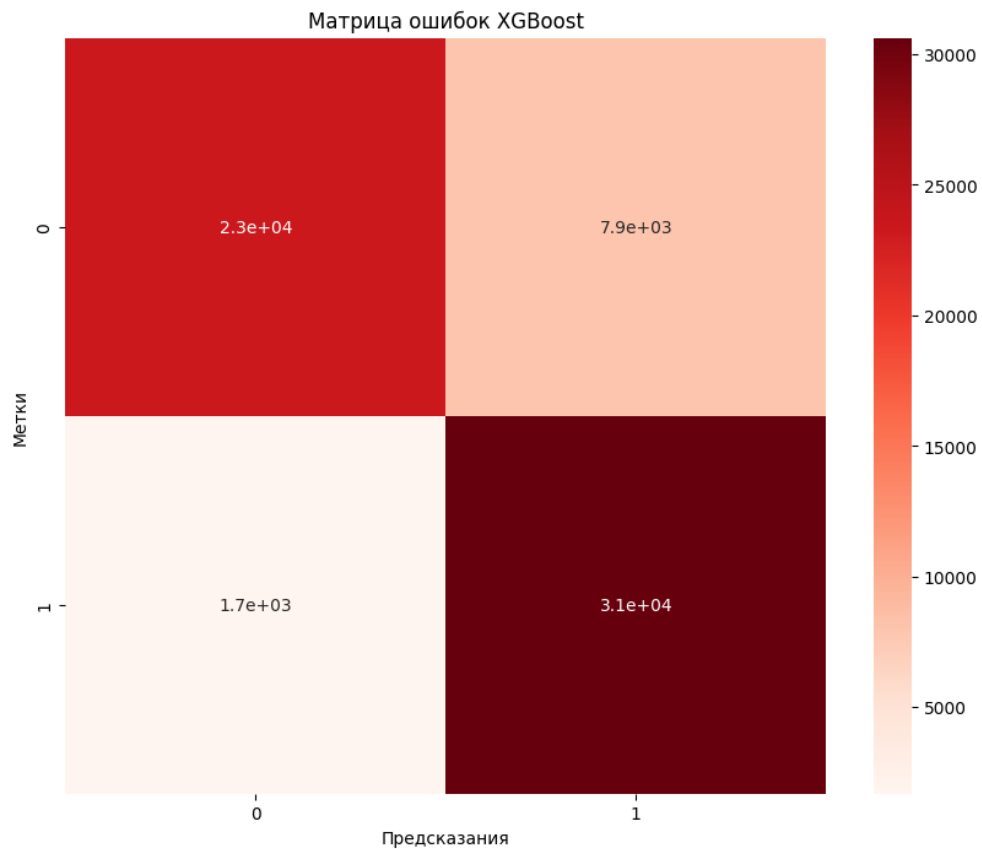
- L1, L2
- Dropout (применительно к NN)

Алгоритмы оптимизации. История обучения MLP

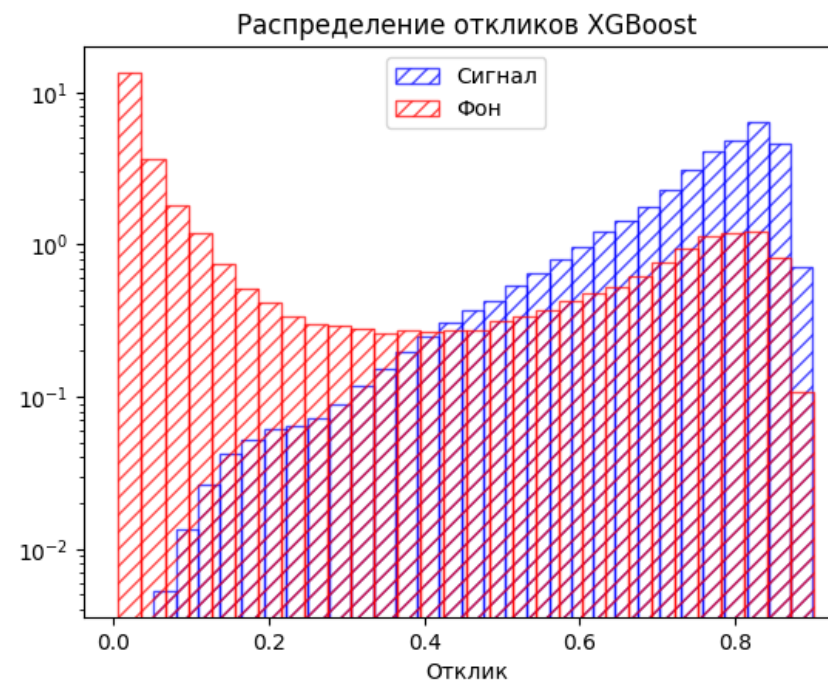
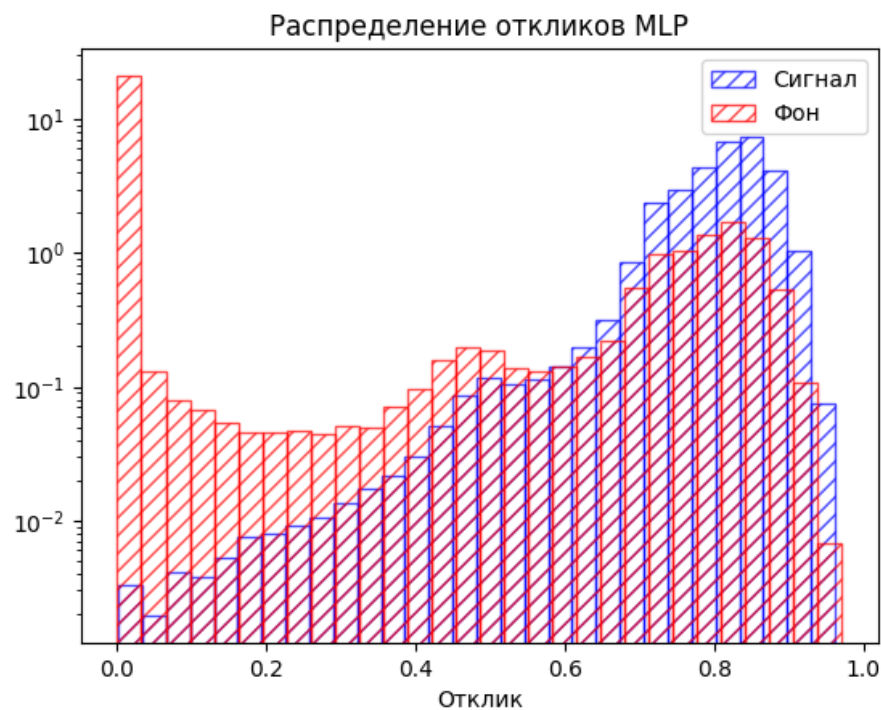
Алгоритм	AUC
Adam	0.849 ± 0.001
SGD	0.799 ± 0.001
RMSProp	0.879 ± 0.001



Матрицы ошибок



Распределение откликов моделей XGBoost и MLP на МК - данных



Формирование признаков.

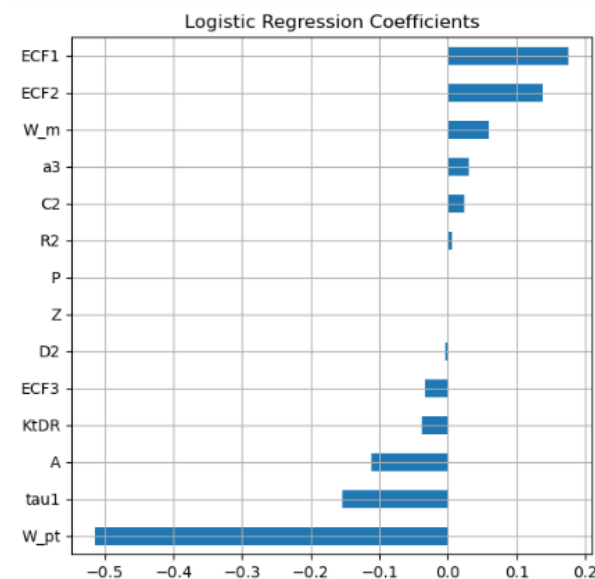
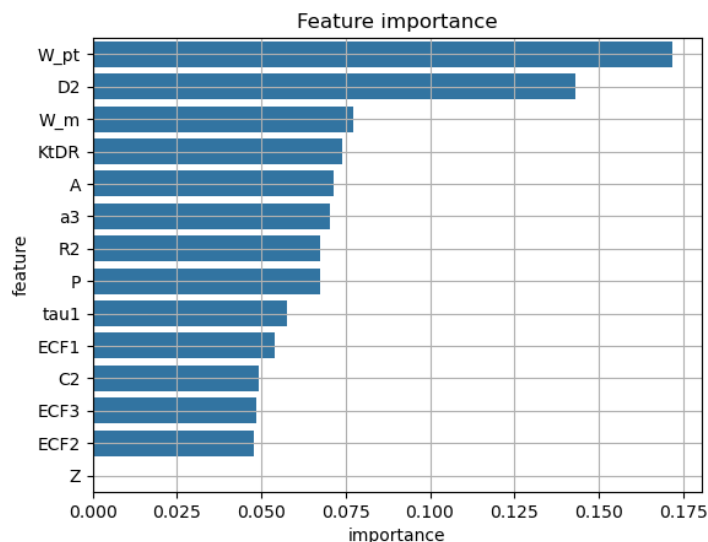
Предварительный анализ



Первично сформированы 14 признаков.

Техники фильтрации:

- Метод с использованием деревьев решений
- Метод с использованием логистической регрессии



По **итогам** исследований для дальнейшего анализа используются 13 переменных, так как при оценке значимости признаков деревом решений и логистической регрессией сделан вывод о том, что переменная Z_{cut} не влияет на дальнейшие результаты моделей.