



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

Научный руководитель:

к.ф.-м.н. Мягков Алексей Григорьевич

Студент:

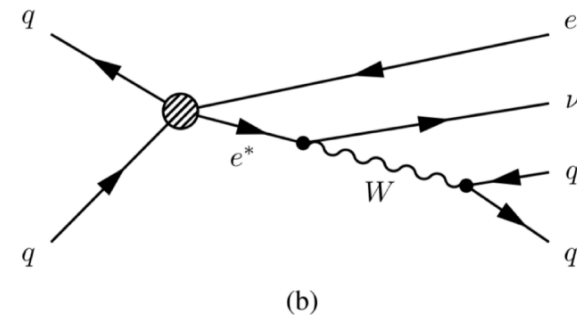
Ван Алина Маошэновна, М23-112

Москва, 2025

Проблемы Стандартной модели

- Скрытая масса
- Проблема иерархии масс и структуры поколений
- Темная энергия и т.д.

Идентификация толстых струй, образованных W -бозонами, распавшимися по адронной моде, является одним из главных составляющих этапов анализа данных с экспериментов по поиску новой физики.



Пример: Поиск возбужденного лептона с последующим распадом через калибровочный бозон

Целью работы является применение методов машинного обучения для решения задачи идентификации событий, образованных W -бозоном.

В соответствии с поставленной целью **задачами** данной работы являлись:

1. Изучение различных алгоритмов машинного обучения с учителем в рамках задачи бинарной классификации;
2. Отбор событий для последующего формирования обучающих и тестовых выборок из Монте-Карло сэмплов с учетом веса каждого события;
3. Предварительный анализ полученных датасетов, формирование признаков и оценка их значимости;
4. Обучение моделей машинного обучения для задачи бинарной классификации: выбор оптимальных гиперпараметров и архитектуры моделей;
5. Тестирование и оценка модели;

Исходные данные. Отборы

Датасет с сигналом строится на основе Монте-Карло модуляций процесса образования пары топ-антитоп.

Отбор осуществлялся по следующим критериям:

- Условие на необходимость как минимум одной b -меченной струи
- Условие на кинематику струи
 - $p_t > 30 \text{ GeV}$
 - $\eta < 2.5$

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

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

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

Датасет с фоном строится на основе Монте-Карло модуляций следующих процессов:

- $V V$ - образование двух векторных бозонов
- $W(\rightarrow e\nu)$ - образование W -бозона с последующим распадом в e и ν
- $Z(\rightarrow ee)$ - образование Z -бозона с последующим распадом в электрон-позитронную пару
- $Z(\rightarrow \tau\tau)$ - образование Z -бозона с последующим распадом в τ -лептонную пару
- $W(\rightarrow \nu\tau)$ - образование W -бозона с последующим распадом в τ -лептон и ν

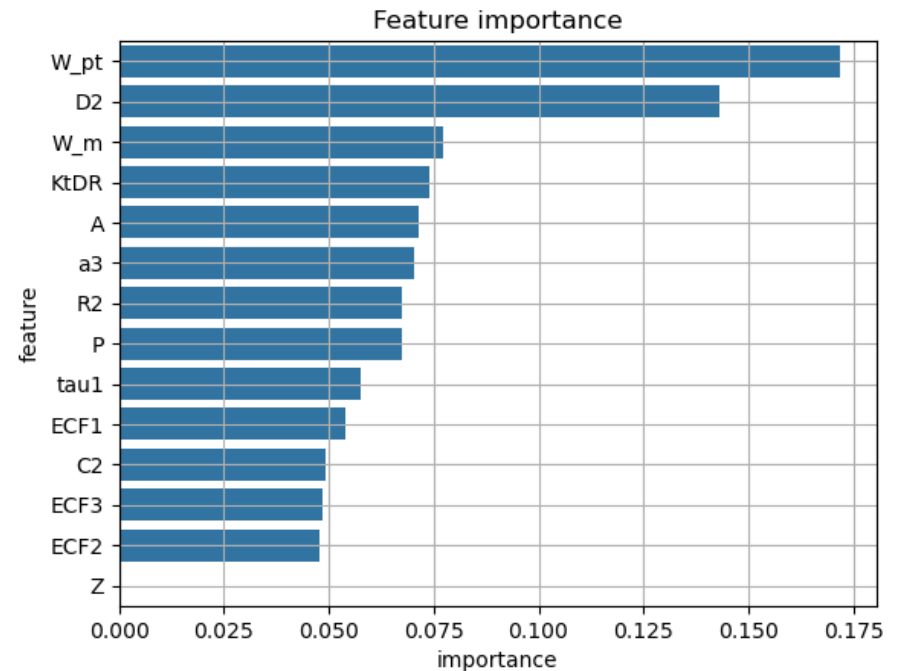
Также в работе используются данные ране 2017 года с эксперимента ATLAS со светимостью 6.1 fb^{-1} .

Формирование признаков. Предварительный анализ



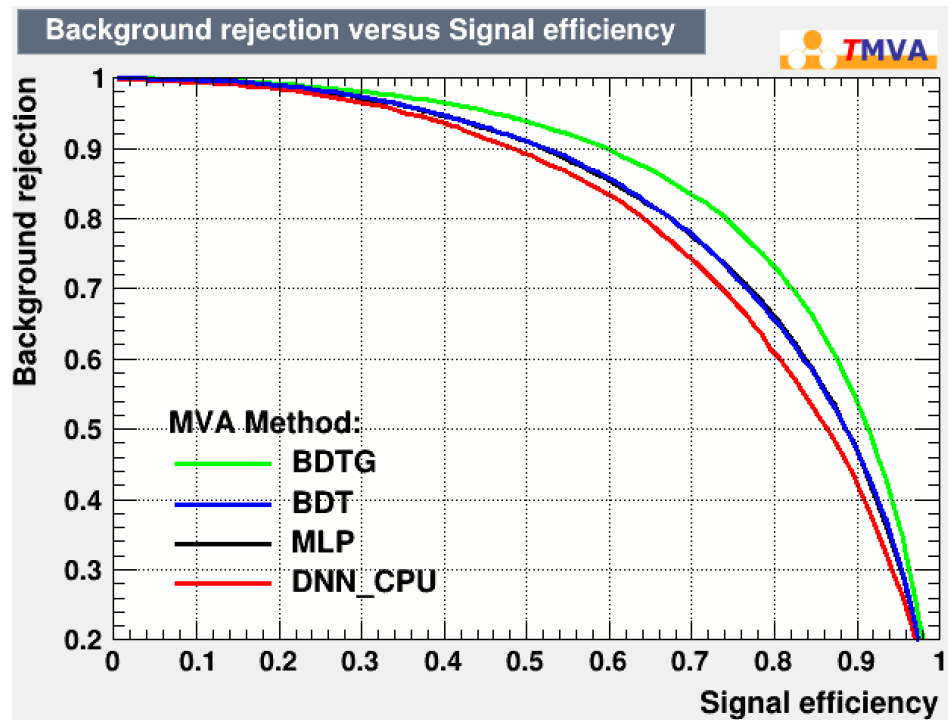
Первично сформированы 14 признаков*. Для устранения весомых внутренних корреляций между признаками и для формирования обучающих/тестовых выборок были применены различные техники фильтрации:

- Методы фильтрации
- Метод обертки
- Метод с использованием деревьев решений
- Метод с использованием логистической регрессии



По итогам исследований выбраны **13 признаков** для обучения модели МО

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)

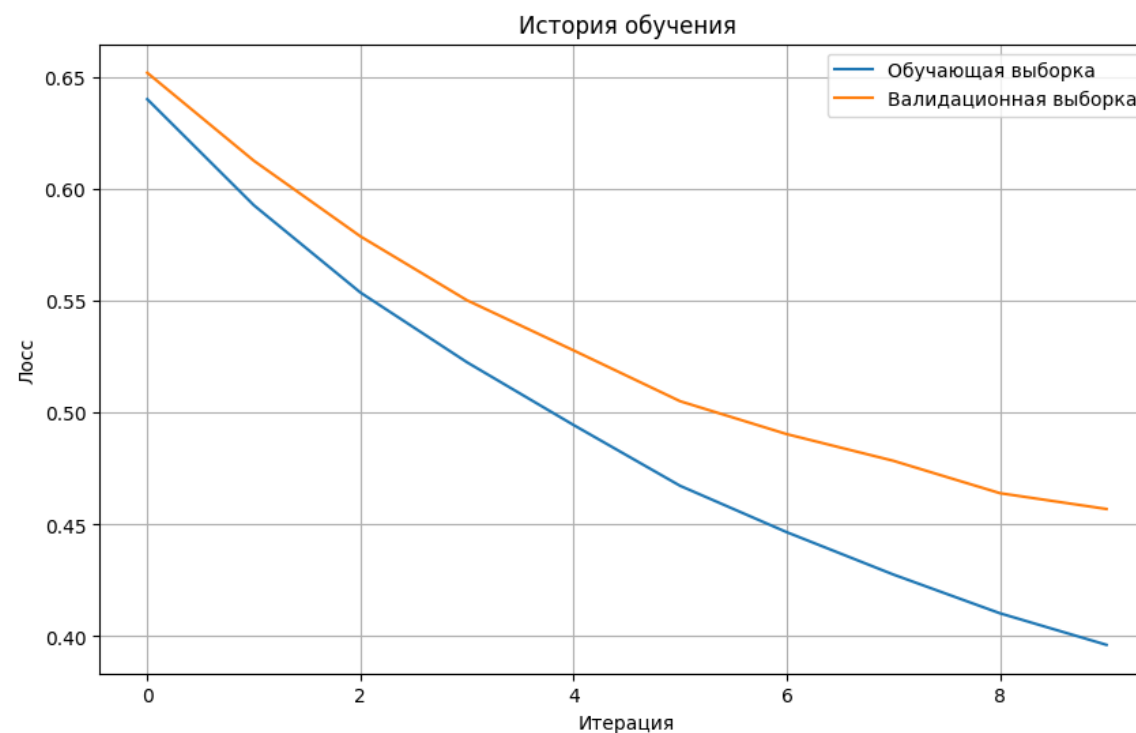
Настройка гиперпараметров

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

Полученные значения гиперпараметров обучаемой модели следующие:

max_depth	8
min_child_weight	3
gamma	0.1
subsample	0.6
colsample_bytree	0.6
learning_rate	0.1
lambda	0.1
alpha	0.1

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



Гиперпараметры

Алгоритм оптимизации: Adam

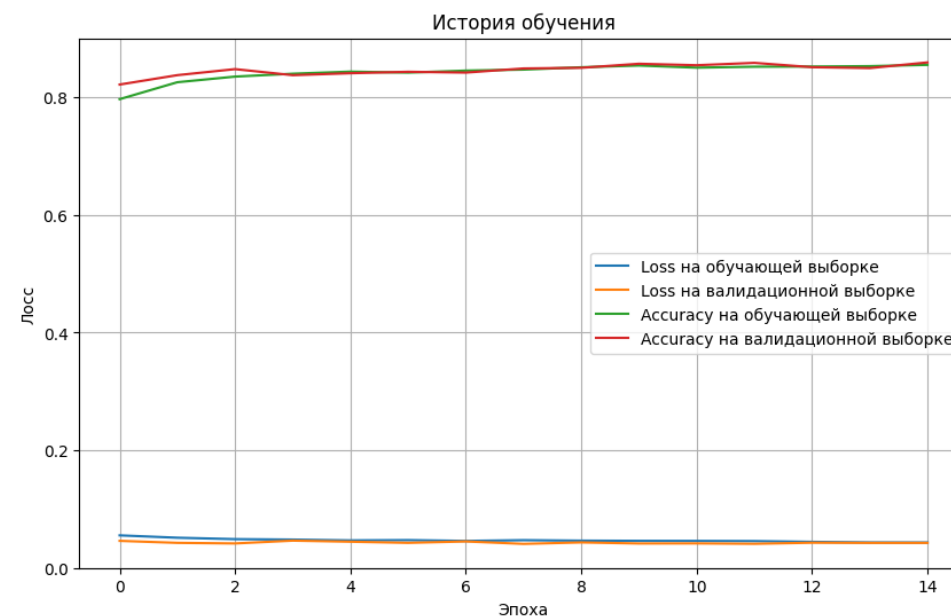
Скорость обучения: 0.001

Размер батча: 128

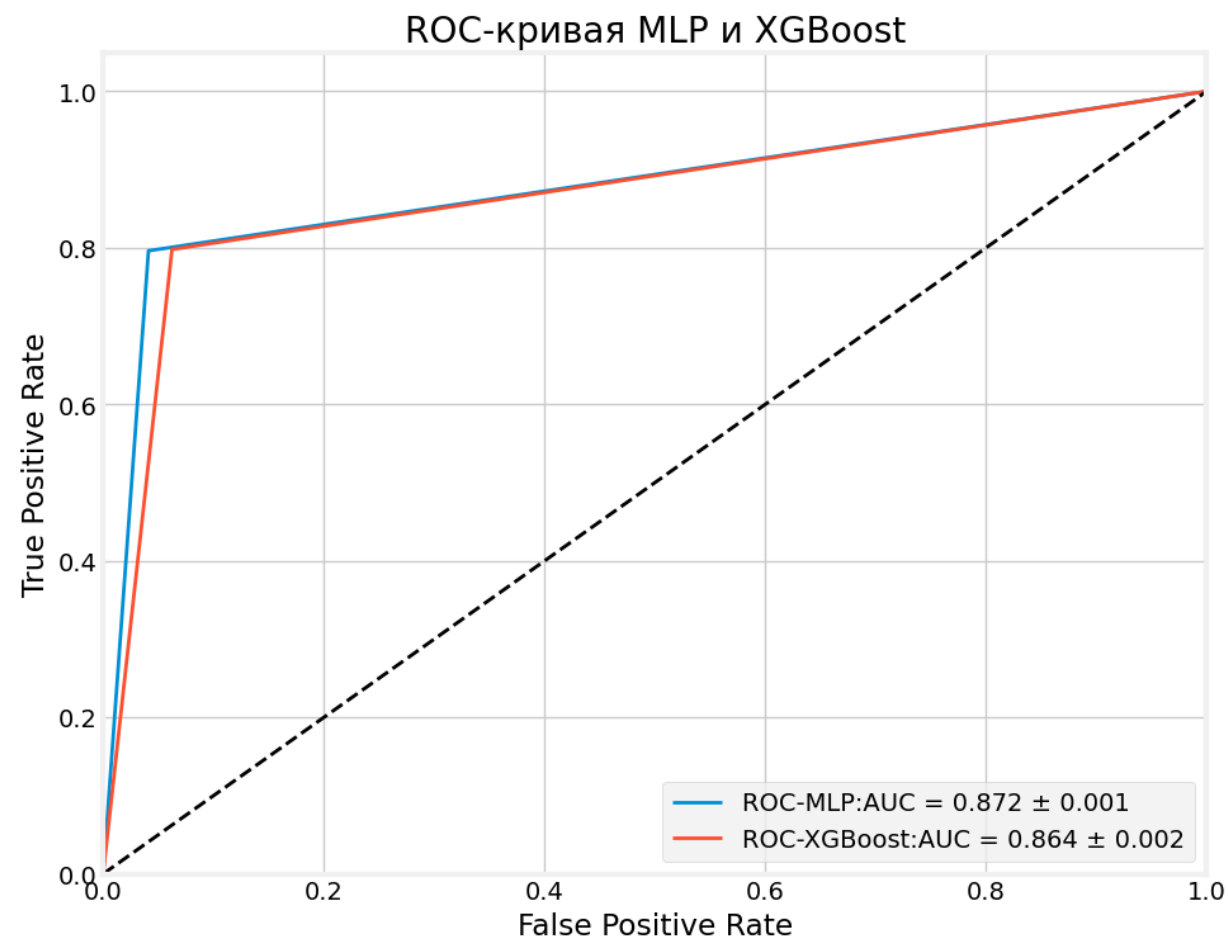
Архитектура

Слой	Нейроны	Функция активации
Входной	13	ReLU
Скрытый слой	32	ReLU
Скрытый слой	64	ReLU
Скрытый слой	128	ReLU
Выходной слой	1	Сигмоида

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



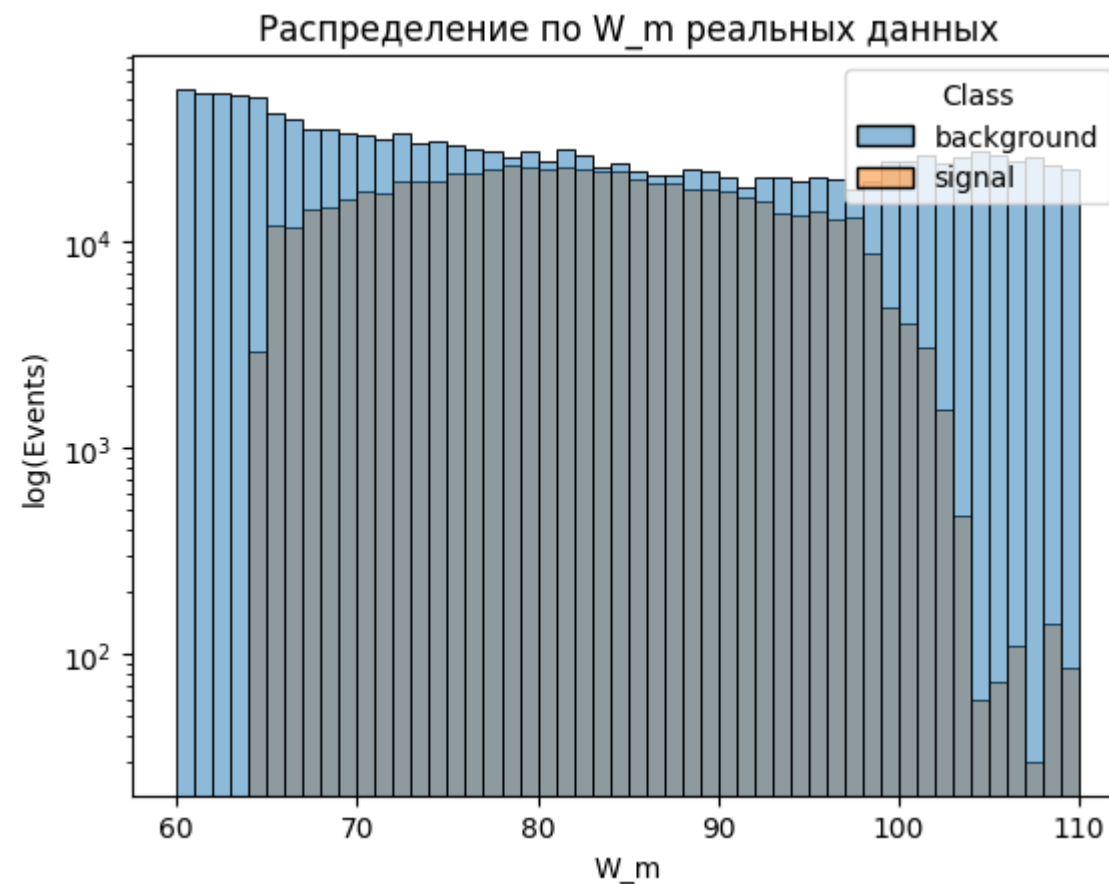
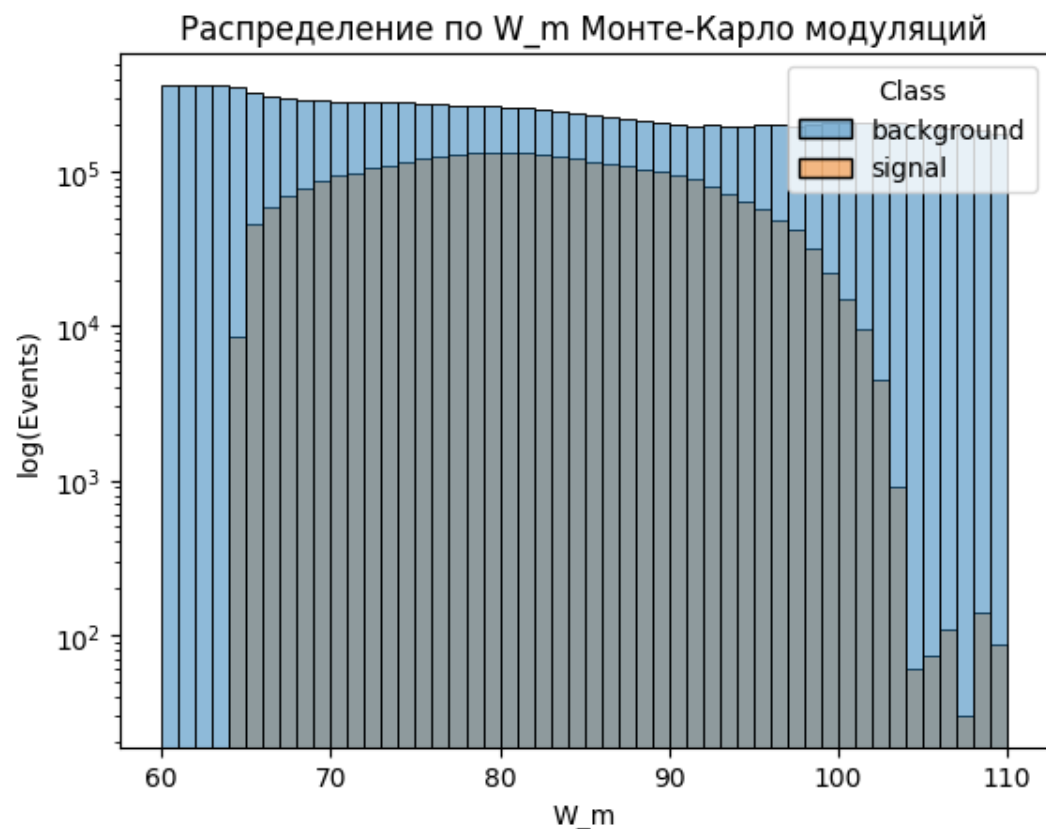
Оценка моделей на тестовых данных



Accuracy MLP = 0.853 ± 0.001

Accuracy XGBoost = 0.849 ± 0.002

Результаты модели XGBoost на реальных данных



Заключение

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

1. Сформированы обучающие и тестовые выборки из Монте-Карло модуляций. Рассчитаны веса событий Монте-Карло модуляций исходя из реальных данных, сформирован датасет с реальными данными (светимость 6.1 fb^{-1}) для последующих исследований отклика обученных моделей
2. Изучены различные алгоритмы МО и техники отбора признаков. В результате изучения данных методов на подвыборке из МК данных выбраны оптимальные признаки и алгоритмы МО
3. Проведено обучение и тестирование MLP и XGBoost моделей с использованием всех фоновых процессов МК модуляций с учетом веса события

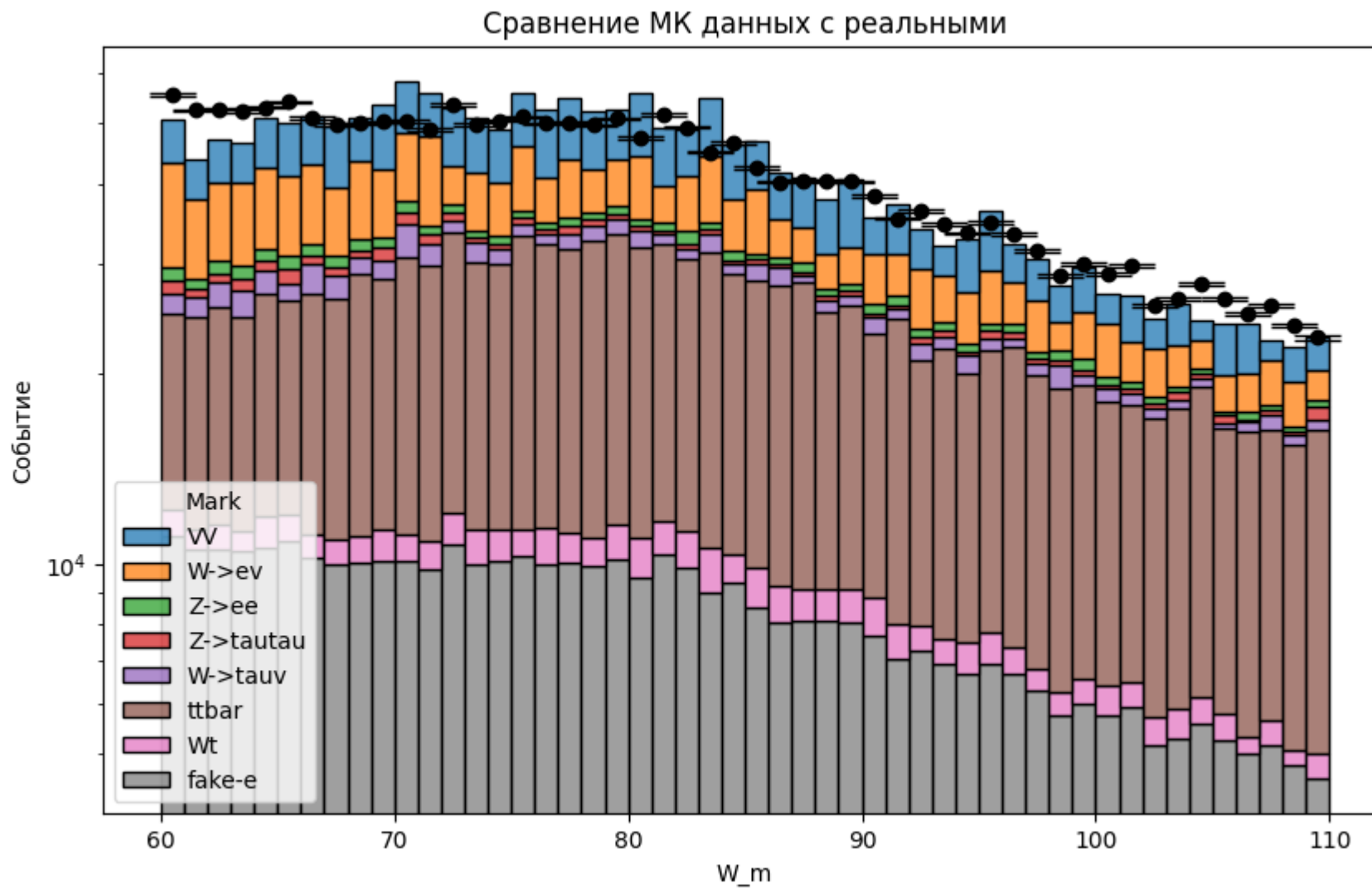
Дальнейшие планы:

Исследование отклика обученных моделей на реальных данных.

Дополнительные слайды



Сравнение Монте-Карло данных и реальных данных



Матрицы ошибок

