МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»
(НИЯУ МИФИ)

ИНСТИТУТ ЯДЕРНОЙ ФИЗИКИ И ТЕХНОЛОГИЙ
КАФЕДРА №40 «ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ»

УДК 531.3, 539.1.05

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ
ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИДЕНТИФИКАЦИИ СТРУЙ, ОБРАЗОВАННЫХ W-БОЗОНОМ

Студент _____ А. М. Ван

Научный руководитель,
к.ф.-м.н. _____ А. Г. Мягков

Москва 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Эксперименты и исследования в физике высоких энергий привели к формированию основной теории строения и взаимодействия частиц. Стандартная модель - это современная теория в физике элементарных частиц, которая описывает сильное, электромагнитное и слабое взаимодействие. Однако, несмотря на все свои преимущества, Стандартная модель не является полной теорией всего. Данная теория не дает описаний таким экспериментальным фактам, как скрытая масса, нейтринные осцилляции, темная энергия, не дает решений проблеме барионной асимметрии и проблеме иерархии масс и структур поколений. Предполагается, что Стандартная модель является частью более общей теории.

Возможность наблюдения новой физики за рамками Стандартной модели является одной из главных задач на LHC. При поиске аномалий в экспериментальных данных идентификация струй, образованных W-бозоном, распавшимся по адронной моде, является одним из главных этапов анализа. К примеру, в эксперименте по поиску возбужденного лептона с конечным состоянием $e\nu J$ процессом, привносящем основной вклад в фон, является образование пары топ-антитоп. Так как топ-кварк распадается преимущественно по слабому взаимодействию, идентификация струй, образованных W-бозоном, является необходимой для подавления фона от данного процесса.

Для идентификации струй, образованных W-бозоном, на данный момент используются различные дискриминирующие и кинематические переменные. При использовании данных методов обрезки есть вероятность неправильной идентификации событий, кинематика которых схожа.

ЦЕЛИ И ЗАДАЧИ

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

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

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

1 ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1 СТАНДАРТНАЯ МОДЕЛЬ

Стандартная модель – теория, описывающая частицы, из которых состоит материя, и их взаимодействие друг с другом. Она появилась в середине 20-го века и приняла свою окончательную форму после всех экспериментальных подтверждений. Модель проясняет обнаруженные к настоящему времени элементарные частицы и их поведение с тремя фундаментальными взаимодействиями: слабым, сильным и электромагнитным. Квантовая хромодинамика дает описание сильному взаимодействию, электрослабая теория описывает слабое и электромагнинтное взаимодействия. Элементарные частицы, определяемые СМ, показаны на рисунке 1.1.

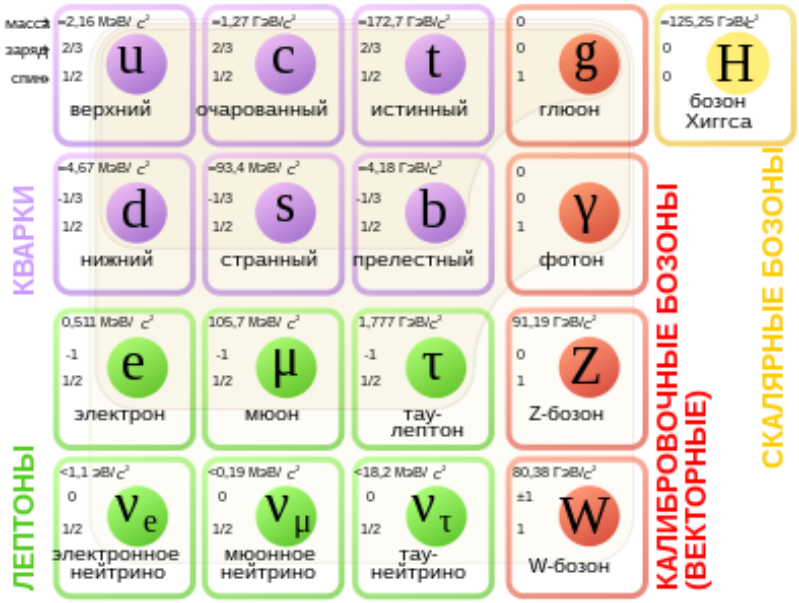

Рисунок 1.1 — Стандартная модель

Стандартная модель содержит фермионы и бозоны. Фермионы [1] в модели подчиняются статистике Ферми-Дирака и имеют спин 1/2. Для опи-

сания фермионов используют биспинорное представление группы Лоренца $SO(3,1)$. Фермионы делятся на две группы: кварки и лептоны. В настоящее время известно 6 лептонов и 6 кварков (по три цвета на каждого) [2], которые, в свою очередь, делятся на три поколения по двое, тем самым образуя дублеты.

Дублеты лептонов удовлетворяют локальной калибровочной симметрии $SU_L(2) \times U_Y(1)$. Электрон, мюон и τ-лептон участвуют в слабом и в электромагнитном взаимодействиях. Их нейтрино участвуют только в слабом взаимодействии. Дублеты кварков удовлетворяют локальной калибровочной симметрии групп $SU(3) \times SU_L(2) \times U_Y(1)$. Кварки участвуют во всех видах взаимодействия.

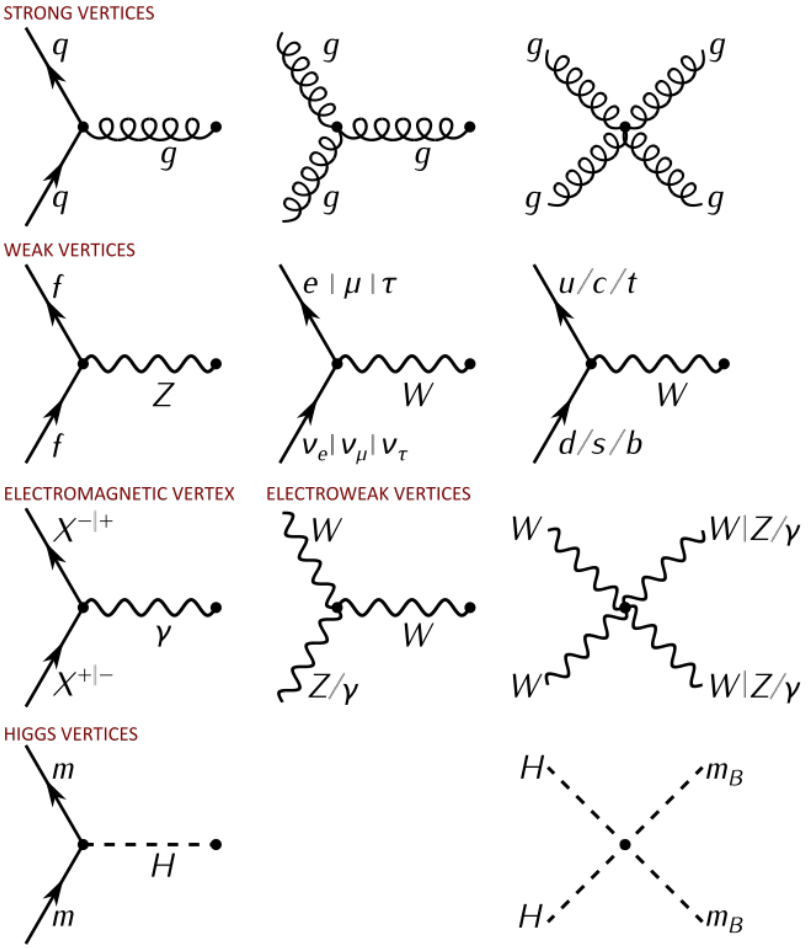

Рисунок 1.2 — Вершины Стандартной модели (m — частица, обладающая массой; m_B — бозон, обладающий массой.)

Бозоны, имеющие целый спин и подчиняющиеся статистке Бозе - Эйнштейна, в свою очередь, являются переносчиками взаимодействия, возникающими вследствие требования локальной калибровочной инвариантности. Частицы W, Z, γ, g имеют спин 1 и являются векторными бозонами. Переносчиками сильного взаимодействия являются 8 глюонов, слабого – $W^{\pm}$, Z-бозоны, электромагнитного – фотоны.

Для придания масс частицам вводится скалярное поле Хиггса. Механизм Хиггса заключается в нарушении симметрии $SU_L(2) \times U_Y(1)$ до $U_{em}(1)$. Квантом поля Хиггса является бозон Хиггса, имеющий спин 0.

На рисунке 1.2 представлены все вершины взаимодействия Стандартной модели.

1.2 МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧЕ БИНАРНОЙ КЛАССИФИКАЦИИ

Машинное обучение[[3]] - это динамично развивающаяся область прикладной математики, которая фокусируется на разработке и применении алгоритмов для решения различных задач с использованием обучающих данных. Машинное обучение позволяет решить задачи, которые недопустимо сложны для традиционных подходов или не имеют точных математических моделей, алгоритмов. При этом в data-driven подходе алгоритм обработки данных и нахождения взаимосвязей между ними заранее не известен и формируется в результате обучения. В рамках данной работы рассматриваются алгоритмы машинного обучения с учителем.

Методы машинного обучения с учителем направлены на изучение взаимосвязи между входными данными (признаками) и выходными данными (для задач классификации метками класса). Постановка задачи для применения алгоритмов машинного обучения следующая:на основе размеченных данных, которые представлены в виде пар [матрица признаков; целевое значение], необходимо найти наилучшую функцию аппроксимации, веса которой минимизируют эмпирический риск. Обучающая выборка используется для настройки внутренних параметров (весов) модели. В зависимости от задачи обучающая выборка также может быть разделена и на валидационную выборку, которая необходима для оценки обобщающих спо-

собностей модели на новых для нее данных. При оценке результатов модели на валидационной выборке в процессе обучения можно вовремя избежать переобучения, т.е. ситуации, в которой модель точно подстраивается под обучающие данные и теряет способность к адекватному прогнозированию. Кроме того, для окончательной оценки результатов модели используется тестовая выборка. Она представляет собой отдельную выборку данных, которые не участвовали в обучении и валидации модели. Это необходимо для объективного оценивания модели на данных, которые модель не видела.

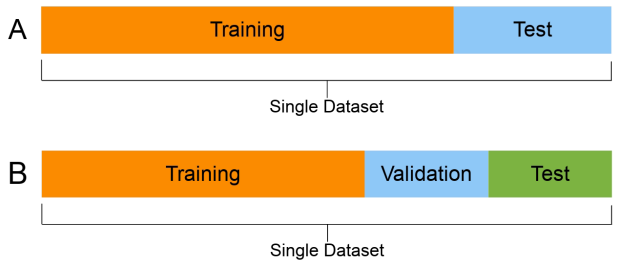

Рисунок 1.3 — Разделение датасета

Основными задачами МО с учителем являются задачи регрессии и задачи классификации. При регрессии алгоритм МО учится предсказывать непрерывное число, в задаче классификации - категориальное значение. В данной работе поставлена задача бинарной классификации, информация о которой будет представлена далее.

1.2.1 БИНАРНАЯ КЛАССИФИКАЦИЯ

В данной работе под постановкой задачи идентификации частицы предлагается постановка задачи бинарной классификации. Отклики обучаемой модели должны быть дискретны и принимать только два значения: 0 (фон) и 1 (сигнал). Сама задача бинарной классификации заключается в нахождении гиперплоскости в признаковом пространстве, которая оптимально разделяет объекты двух классов, т.е. в данной задаче события от струй, образованных W-бозоном, и события от фоновых процессов.

Функция потерь в контексте бинарной классификации - это мера неточности предсказания обучаемого классификатора для объектов выборки. Она количественно оценивает разницу между фактическими метками класса и предсказанными откликами модели. Алгоритмы машинного

обучения направлены на минимизацию функции потерь для нахождения оптимальных внутренних параметров классификатора и, как следствие, улучшения производительности модели на новых данных.

Типичной функцией потери в бинарной классификации является бинарная кросс-энтропия. Формально, функцию потерь для одного объекта выборки можно записать как:

$$L(y, \hat{y}) = -\alpha \left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \right], \tag{1.1}$$

где y — истинная метка класса, $\hat{y}$ — предсказанная вероятность принадлежности классу, α — вес события.

Для регрессии функция потерь является одной из метрик, по которой можно охарактеризовать качество и производительность модели, так как в регрессии отклики принимают непрерывные значения. В случае бинарной классификации оценка модели по среднему значению функций потерь не достаточна, так как отклики модели дискретны. Поэтому необходимы другие механизмы оценки работы модели.

Точность (Accuracy)

Данная метрика определяет долю правильно классифицированных событий среди всех событий. Более формально данную метрику можно представить в виде:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}, \tag{1.2}$$

где

- TP — сумма весов событий, принадлежащих положительному классу и предсказанных как положительный класс;
- TN — сумма весов событий, принадлежащих отрицательному классу и предсказанных как отрицательный класс;
- FN - сумма весов событий, которых классификатор ложно отнес к отрицательному классу;
- FP - сумма весов событий, которых классификатор ложно отнес к положительному классу.

ROC-кривая и AUC

Для оценки работы бинарного классификатора также используют ROC-кривую. При различных порогах классификации (от 0 до 1) она отображает соотношения между TPR и FPR:

$$TPR = \frac{TP}{TP + FN}, \tag{1.3}$$

$$FPR = \frac{FP}{FP + TN}. \tag{1.4}$$

AUC - это интеграл под ROC-кривой. Значение данной метрики варьируется между 0 и 1, где 1 указывает на идеальную классификацию. Худшим значением AUC является значение, равное 0.5. Оно означает, что классификатор работает, как случайное подбрасывание монетки. AUC и ROC-кривая полезны тем, что позволяют оценить общую способность модели и учитывают все возможные пороги классификации.

Для решения задачи бинарной классификации существует множество алгоритмов: от простых моделей, как логистическая регрессия, до глубоких нейронных сетей. В работе использованы алгоритмы MLP и BDT, речь о которых пойдет далее.

1.2.2 BDT

Одним из базовых подходов к решению задачи бинарной классификации является алгоритм *дерева решений*. Идея данного метода заключается в разбиении признакового пространства на различные множества непересекающихся областей и дальнейшей постановке этим областям меток класса. Для данного процесса разбиения необходимо правило, следуя которому объект выборки попадает в определенную область. Дерево решений - совокупность этих правил.

Правило ветвления в дереве решений определяется так, чтобы среди различных признаков выбрать тот, условие по которому будет сокращать неоднородность в секторах признакового пространства. Количественная мера неоднородности в дереве решений может определяться индексом Джини:

$$G(R_l) = \sum_{k=1}^{K} p_{lk}(1 - p_{lk}), \tag{1.5}$$

где p_{lk} - это вероятность k-го класса в области R_l, K - это количество меток класса. То есть, когда индекс Джини равен 0, можно сделать вывод, что данная область содержит объекты только одного класса.

Критерий останова итераций разбиения на ноды определяется в зависимости от задачи. Популярные из них: ограничение на количество листьев, на глубину дерева и т.д.

По смыслу своего формирования дерево решений при его углублении склонно к переобучению. Чем больше глубина дерева, тем более вероятно то, что модель потеряет обобщающие способности и будет следовать за выбросами. Поэтому часто эту базовую модель используют в *ансамблевых методах.*

Бустинг — это метод ансамблевого обучения, который последовательно соединяет несколько слабых моделей для создания более сильной модели. Он направлен на обучение новых моделей для исправления ошибок, допущенных предыдущими. Тем самым данный процесс повышает общую работоспособность модели.

При градиентном бустинге каждая новая модель обучается для минимизации функции потерь предыдущей с помощью градиентного спуска. На каждой итерации алгоритм вычисляет градиент функции потерь по отношению к прогнозам, а затем обучает новую слабую модель для минимизации этого градиента. Потом прогнозы новой модели добавляются в ансамбль, и обучение повторяется до тех пор, пока не будет достигнут критерий останова.

1.2.3 MLP

Нейронная сеть - это сложная дифференцируемая функция, которая преобразует пространство признаков в пространство откликов. Все ее внутренние веса связаны между собой и настраиваются одновременно (когда мы их специально не "замораживаем"). Конкретно данную сложную функцию представляют в виде композиции простых функций, которые называют слоями нейронной сети. Количество данных слоев в сети определяет уровень сложности и глубину модели [[4]].

Многослойный персептрон (MLP) - это один из типов нейронной сети, который состоит исключительно из полносвязных слоев. Архитектура MLP довольно проста: в ней содержатся входной слой, в котором обязательно должно содержаться количество нейронов, равное размерности признакового пространства, выходной слой и скрытые слои [[5]].

Полносвязные слои - слои, в которых выходные нейроны связаны со всеми входными нейронами. Выходные данные полносвязного слоя формально можно представить с помощью следующей формулы:

$$y = \sigma(Wx + b), \tag{1.6}$$

где y - выход слоя, x — вектор входных данных, поступающих из предыдущего слоя, W — матрица весов, связывающая вход и выход нейрона, b — интерсепт, σ — функция активации.

Функция активации - это нелинейное преображение, которое поэлементно применяется к пришедшим на вход данным. Благодаря данной нелинейности, нейронные сети способны порождать более информативные признаковые пространства, в которых функция ошибки, которую мы пытаемся минимизировать, будет оптимальной. Главные требования для функции активации: быть монотонной и иметь первую производную почти всюду (необходимо для обратного распространения ошибки).

В качестве функции активации могут использоваться разные нелинейные функции, и у каждой из них есть свои плюсы и минусы. Одни из часто используемых функций активации - это ReLu, сигмоида и гиперболический тангенс.

Ключевые механизмы обучения MLP:

1) Прямой проход (forward pass): При прямом распространении данные поступают от входного слоя к выходному, проходя через все скрытые слои. Нейрон вычисляет взвешенную сумму входных, после чего эта сумма проходит через функцию активации для введения нелинейности.
2) Функция потерь: Следующим шагом после выявления отклика модели является вычисление взвешенной функции потерь, которая сравнивает отклики модели с фактическими метками. Для бинарной классификации формульное представление функции потерь представлено выше [1.1].
3) Обратное распространение ошибки (back pass): На данном этапе вычисляется градиент функции потерь по каждому весу модели (в том числе интерсепту) согласно правилу дифференцирования сложной функции. Нейронная сеть обновляет весовые коэффициенты и интерсепт, двигаясь в направлении, противоположном градиенту, чтобы уменьшить потери.

Существует множество модификаций алгоритма градиентного спуска. Популярные из них: SGD, Adam, RMSprop. В данной работе используется алгоритм оптимизации обучения Adam. Одним из немаловажных положительных факторов данного алгоритма является то, что вероятность застревания модели в локальном минимуме уменьшается. Это происходит за счет того, что веса в Adam с каждой итерацией обновляются экспоненциальным скользящим средним прошлых градиентов, что устраняет незначительные колебания и позволяет алгоритму быстрее сходиться.

2 ДЕТЕКТОР ATLAS

Детектор ATLAS [6] предназначен для исследования широкого спектра физических явлений: от измерения свойств частиц Стандартной модели до поиска новой физики за ее рамками. Он используется для детектирования и идентификации частиц. Детектор ATLAS представляет собой многоцелевой 4π-детектор с симметричной цилиндрической геометрией. Установка состоит из серии больших концентрических цилиндров вокруг линии пучка. Детектор ATLAS состоит из внутреннего трекового детектора, электромагнитного и адронного калориметров, мюонного спектрометра и магнитных систем. Каждый из них в свою очередь сделан из повторяющихся слоев. Трековый детектор предназначен для определения параметров треков заряженных частиц для измерения их импульса. Калориметры необходимы для измерения энерговыделения частиц, мюонная система используется для определения импульса и направления пролёта высокопроникающих мюонов. Магнитная система предназначена для искривления траекторий заряженных частиц для определения их импульса. Устройство детектора ATLAS представлено на рисунке 2.1.

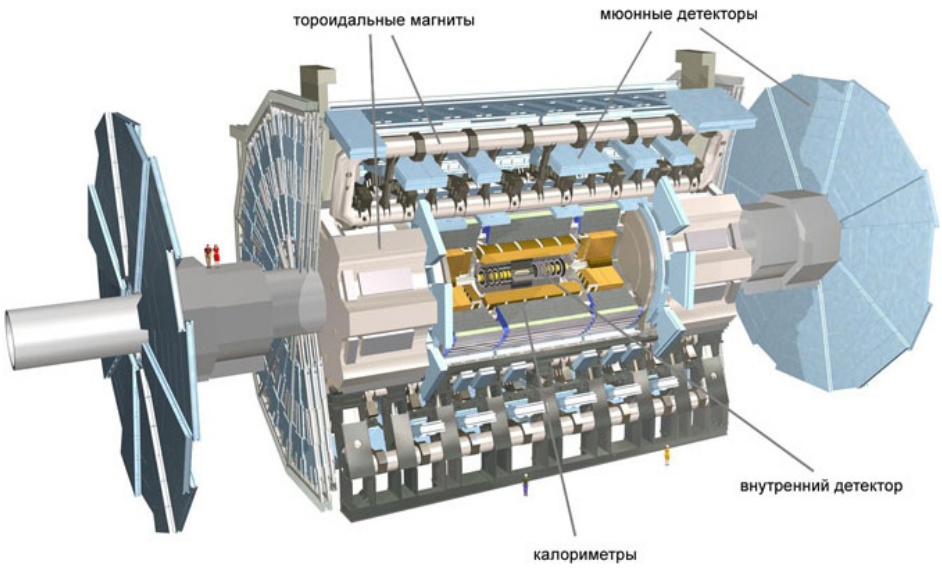

Рисунок 2.1 — Устройство детектора ATLAS

2.1 ПРОТОН-ПРОТОННОЕ СТОЛКНОВЕНИЕ

Протон – это барион, представляющий собой структурную частицу: он состоит из трех кварков uud, скрепленных вместе глюонным полем. В партонной модели при высоких энергиях протон рассматривается как ансамбль невзаимодействующих партонов: глюонов и кварков. Столкновение с жестким рассеянием можно рассматривать как взаимодействие между двумя партонами каждого протона, каждый из которых несет долю импульса x1, x2 взаимодействующих протонов. При столкновении двух партонов происходит жесткий процесс, описываемый СМ, образовавшиеся кварки и глюоны переходят в бесцветные адроны в процессе адронизации. Энергия столкновения расходуется на рождение большого числа адронов.

2.2 КИНЕМАТИКА LHC

В детекторе ATLAS используется несколько основных систем отсчета. Начало отсчета выбирается в точке взаимодействия, ось x расположена к центру LHC, ось z направлена вдоль движения пучка, ось y направлена вверх. В цилиндрической системе координат полярный угол θ отсчитывается от положительного направления оси z , азимутальный угол ϕ определяется в плоскости Oxy вокруг оси пучка.

Кинематика объектов событий описывается следующими переменными:

- Из-за того, что распределение частиц не изотропно, а прижато к осям, вместо угла θ используется псевдобыстрота
- Поперечный импульс p_t
- Энергия E
- Азимутальный угол ϕ

3 ИСХОДНЫЕ ДАННЫЕ

Обучение моделей проводится с данными, полученными методом Монте-Карло моделирования протон-протонного столкновения в детекторе ATLAS на LHC с энергией в системе центра масс 13 ТэВ для конечного состояния evJ. В качестве сигнальных событий используются смоделированные сэмплы процесса образования пары $t\bar{t}$. Данный процесс моделировался с помощью МК генераторов Powheg + Pythia 8. В качестве фоновых событий для обучения использовались следующие процессы:

- $Single - t$ - образование одиночного топ-кварка
- VV - 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 e^- и e^+
- $Z(\rightarrow \tau\tau)$ - образование Z-бозона с последующим распадом в τ^- и τ^+
- $W(\rightarrow \nu\tau)$ - образование W-бозона с последующим распадом в τ и ν

Отклик моделей будет исследоваться на экспериментальных данных с RUN2.139 2017 года с энергией в системе центра масс 13 ТэВ общей светимостью $L = 6.4fb^{-1}$.

3.1 ДИСКРИМИНИРУЮЩИЕ И КИНЕМАТИЧЕСКИЕ ПЕРЕМЕННЫЕ

Так как масса W-бозона больше массы типичной КХД-струи, масса струи является основным наблюдаемым показателем, отличающим толстую струю от КХД-струи. Поэтому важно, чтобы в обучающей выборке модели одним из признаков была инвариантная масса [[7]].

Основная масса толстой струи, образованной W-бозоном, обусловлена кинематикой двух составляющих струи, которые соответствуют двум распадающимся кваркам. Для кинематики продуктов распада W-бозона

по адронной моде характерно наличие достаточно большого поперечного импульса. По этой причине поперечный импульс струи также должен характеризовать объект выборки для модели[[8]].

Также в работе используются дискриминирующие переменные[[9]], о которых будет написано далее. Определение переменной D_2 следующее:

$$D_2 = E_{CF3} \times \Big(\frac{E_{CF1}}{E_{CF2}}\Big)^3, \tag{3.1}$$

где энергетические корреляционные функции задаются формулами:

$$E_{CF1} = \sum_{i}^{n} p_{T,i}; E_{CF2} = \sum_{i,j}^{n} p_{T,i} p_{T,j} \Delta R_{ij}; E_{CF3} = \sum_{i,j,k}^{n} p_{T,i} p_{T,j} p_{T,k} \Delta R_{ij} \Delta R_{jk} \Delta R_{ki}. \tag{3.2}$$

ΔR_{ij} определяется как:

$$\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2. \tag{3.3}$$

Также иногда вместо переменной D2 для выделения толстой струи используют переменную C2, которая задана следующей формулой:

$$C_2 = \frac{E_{CF1} \times E_{CF2}}{E_{CF3}^2} \tag{3.4}$$

Также в работе используются переменные, обусловленые геометрическими характеристиками толстой струи:

1) Энергетические корреляционные функции $D2, C2, ECF2, ECF3$
2) Planar flow (измеряет степень, в которой энергия струи равномерно распределяется по плоскости поперек поверхности струи по сравнению с линейной распределением по поверхности струи)
3) N-subjettiness (для эффективного подсчета подструй в струе)
4) Апланарность A
5) Момент Фокса-Вольфрама R_2
6) Angularity (переменная, чувствительная к степени симметрии потока энергии внутри струи)
7) KtDR

4 ПРОЦЕСС РАБОТЫ

4.1 ПОДГОТОВКА ДАННЫХ

Сигнальный датасет строится на данных фонового процесса образования пары топ-антитоп, так как топ преимущественно распадается через W-бозон. Ограничения, поставленные на сигнальное дерево, схожи с контрольной областью пары топ-антитоп: отбираются события с как минимум одной b-меченной струей, поперечный импульс которой больше 30 ГэВ и модуль псевдобыстроты меньше 2.5.

Датасет с фоновыми событиями строится на МК данных остальных вышеописанных процессов. Также поставлены ограничения на отсутствие b-меченных струй в событиях данного дерева.

И на фоновую, и на сигнальную выборку наложены ограничения на поперечный импульс толстой струи и на ее инвариантную массу. Ограничения представлены ниже:

$$60\text{GeV} < m < 110\text{GeV}, p_t > 200\text{GeV}$$

Для некоторых событий вычисление ECF3 не возможно. Переменные D2, C2 задавались по формулам, поэтому для событий с ECF3 равным 0 значения этих переменных были не определены. Для дальнейшей работы с данными и для последующего обучения моделей необходимо было удалить события, содержащие неопределенные значения.

Также были расчитаны датасеты с реальными данными. Гистограмма распределения по инвариантной массе струи для реальных данных и всех процессов МК-модуляций представлена далее.

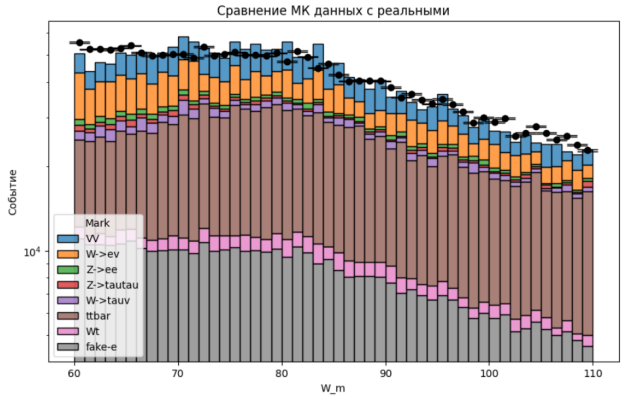

Рисунок 4.1 — Сравнение МК-модуляций и реальных данных

4.2 ОТБОР ПРИЗНАКОВ

Метод фильтрации: Корреляция Пирсона

Метод фильтрации с использованием корреляции Пирсона в машинном обучении применяется для оценки зависимостей между переменными. Этот метод позволяет выявить, какие признаки имеют значительное влияние на целевую переменную. Корреляция Пирсона измеряет степень и направление линейной связи между двумя переменными, принимая значения от -1 до 1. Значение, близкое к 1, указывает на сильную положительную корреляцию; значение, близкое к -1, указывает на сильную отрицательную корреляцию. Использование корреляции Пирсона позволяет отбирать наиболее значимые признаки, что способствует улучшению производительности модели и снижению риска переобучения, как следствие из устранения проблемы мультиколлинеарности.

Для полученных данных ниже представлена матрица корреляции для 13 признаков и целевой переменной.

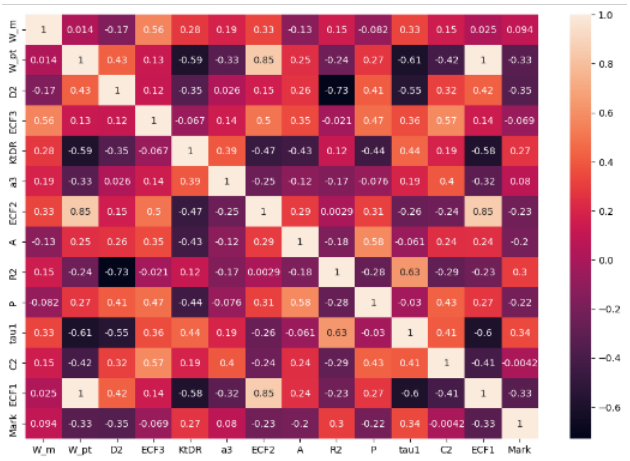

Рисунок 4.2 — Корреляционная матрица

Метод отбора признаков с использованием дерева решений

Основная идея деревьев решений заключается в том, чтобы разбить данные на подмножества, основываясь на значениях признаков. Каждый узел дерева представляет собой условие на определенный признак, а ветви — возможные исходы этого условия. В конечных узлах находятся предсказания модели. Важность признаков может быть оценена на основе критериев, таких как уменьшение энтропии или индекс Джини, что позволяет количественно оценить вклад каждого признака в модель.

Для оценки важности признаков вся полученная выборка проходила через дерево решений с критерием индекса Джини. Получена важность данных признаков, представленная ниже.

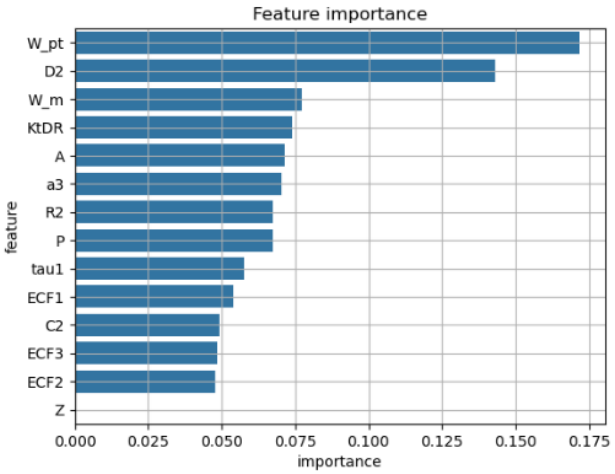

Рисунок 4.3 — Важность признаков

Метод отбора признаков с использованием логистической регрессии и регуляризации L1

Логистическая регрессия находит взаимосвязь между целевой переменной и одной или несколькими независимыми переменными, оценивая вероятности с помощью своей логит-функции. Линейные модели работают так, что коэффициенты при признаках указывают насколько сильно изменение признака влияет на целевую переменную. При большом признаковом пространстве для уменьшения риска переобучения используют методы регуляризации. В частности Lasso регуляризация, а именно прибавление к функции потерь суммы абсолютных значений коэффициентов, позволяет обнулять коэффициенты неинформативных признаков.

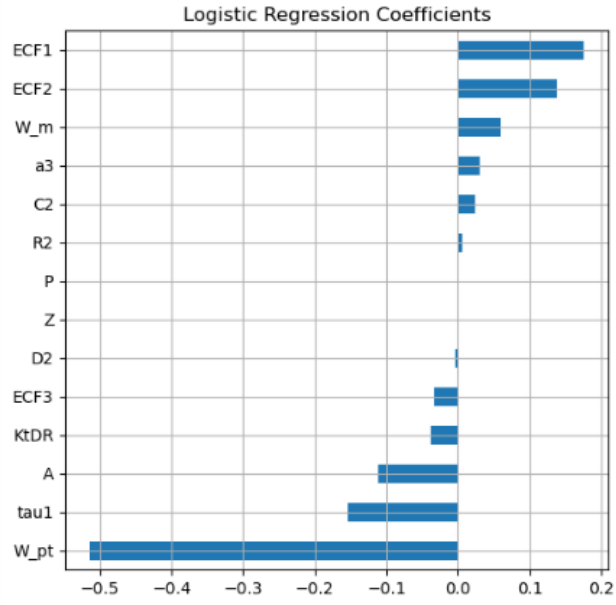

Рисунок 4.4 — Коэффициенты логистической регрессии при независимых переменных

Метод обертки: последовательный отбор признаков

Метод обертки - это процесс выбора признаков, основанный на конкретном алгоритме машинного обучения, который мы используем. Он следует подходу жадного поиска, оценивая все возможные комбинации признаков по определенному критерию. Методы оболочки обычно обеспечивают лучшую точность прогнозирования чем методы фильтрации. Одан-

ко данные методы самые вычислительно затратные из всех, что описывались выше. В последовательном отборе признаков мы начинаем со всех доступных характеристик датасета и строим на их основе модель. Затем мы удаляем переменную из модели, которая дает наихудшее значение меры оценки. Этот процесс продолжается до тех пор, пока не будет достигнут заданный критерий.

Для первой итерации признаков были получены следующие значения важности признаков.

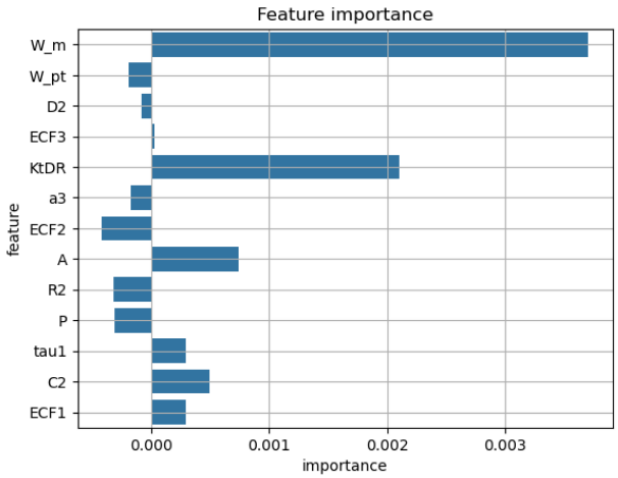

Рисунок 4.5 — Важность признаков в MLP

Вывод

По итогам исследований результатов применения техник отбора признаков выбраны 13 переменных для обучения моделей МО.

4.3 ОБУЧЕНИЕ МОДЕЛЕЙ

4.3.1 XGBOOST

Процесс обучения бинарного классификатора XGBoost включает несколько ключевых этапов. Сначала данные разделяются на обучающую и тестовую выборки, после чего на обучающей выборке создается модель, которая последовательно строит деревья решений, минимизируя ошибку предсказания. В процессе обучения используются методы регуляризации и оптимиза-

ции, что позволяет улучшить обобщающую способность модели и избежать переобучения.

Настройка гиперпараметров в XGBoost с помощью алгоритма жадного поиска включает в себя систематический подход к выбору оптимальных значений параметров модели. Поиск по сетке позволяет протестировать различные комбинации гиперпараметров, таких как max_depth, learning_rate, n_estimators и других, чтобы найти наилучшие настройки для повышения производительности модели.

В работе были определены настраиваемые гиперпараметры и их допустимые диапазоны. Алгоритм жадного поиска был настроен на 5-кратную перекрестную проверку моделей. Модели обучались на МК-данных с учетом веса событий. Диапазон циклов обучения конкретно одной модели был ограничен 10 итерациями.

Так как ансамбли деревьев с градиентным бустингом склонны к переобучению, одними из настраиваемых гиперпараметров были коэффициенты L1 и L2 регуляризации. Следует отметить, что оптимальные гиперпараметры выбирались исходя из значения метрики AUC. Оптимальные гиперпараметры для данной модели представлены далее.

Параметр	**Значение**
objective	binary:logistic
eval_metric	logloss
tree_method	hist
max_depth	8
min_child_weight	3
gamma	0.1
subsample	0.6
colsample_bytree	0.6
learning_rate	0.1
sampling_method	uniform
grow_policy	lossguide
lambda	0.1
alpha	0.1

Таблица 4.1 — Гиперпараметры модели XGBoost для задачи бинарной классификации

Обучение модели происходило согласно найденным гиперпараметрам. В процессе обучения при превышении 10 итераций функция ошибки

на валидационной выборке стала выходить на плато. По причине этого обучение остановлено на 10 итерации. История обучения представлена далее.

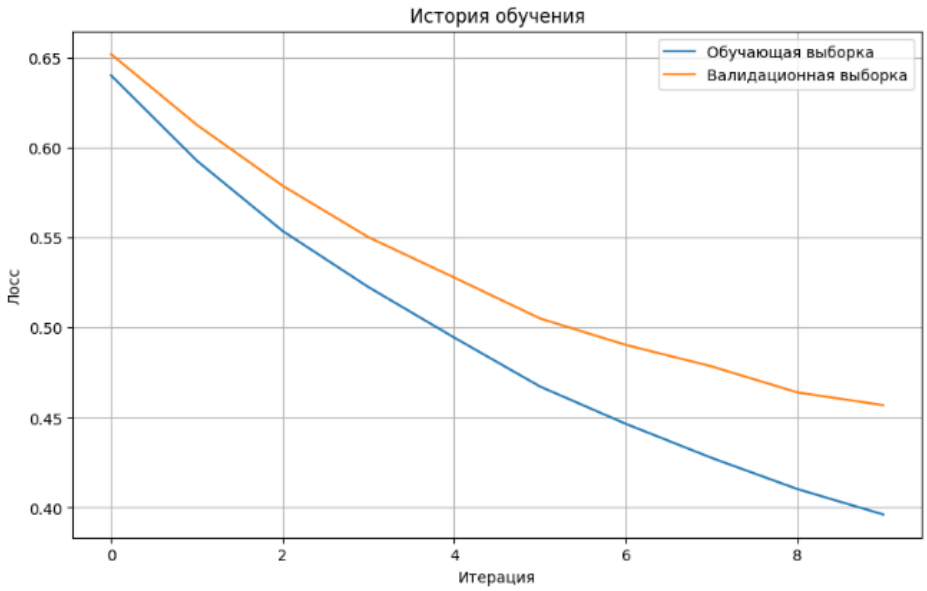

Рисунок 4.6 — История обучения XGBoost

4.3.2 MLP

Обучение полносвязной нейронной сети включает в себя процесс настройки весов нейронов с помощью алгоритма обратного распространения ошибки. Этот метод позволяет минимизировать разницу между предсказанными и реальными значениями, что достигается через итеративное обновление весов на основе градиентного спуска.

Для большого объема обучающих данных реализация алгоритма градиентного спуска на прямую вычислительно затратна и не оптимальна. Однако существует много алгоритмов оптимизации градиентного спуска. В данной работе такой модификацией выбран алгоритм Adam, так как он позволяет обновлять веса нейронов динамически и имеет способность сходиться быстрее.

Следует отметить, что все выборки для обучения, валидации и тестирования предварительно были нормализованны.

Скорость обучения и батч подбирались вручную. Скорость обучения для компиляции данной модели составляет 0.001, батч равен 128. Исходя из опыта предыдущих работ основной функцией активации была выбрана ReLU, так как при использовании данной функции вероятность затухания градиента становится меньше. Архитектура данной сети выбрана следующая:

1) Входной слой с 13 нейронами (в соответствии с 13 признаками);
2) Первый скрытый слой с 32 нейронами и функцией активацией ReLU;
3) Второй скрытый слой с 64 нейронами и функцией активацией ReLU;
4) Третий скрытый слой с 128 нейронами и функцией активацией ReLU;
5) Выходной слой с 1 нейроном и сигмоидой;

В процессе обучения при превышении 10 эпох валидационный лосс выходит на плато. По причине этого обучение остановлено на 10 эпохе. История обучения представлена далее.

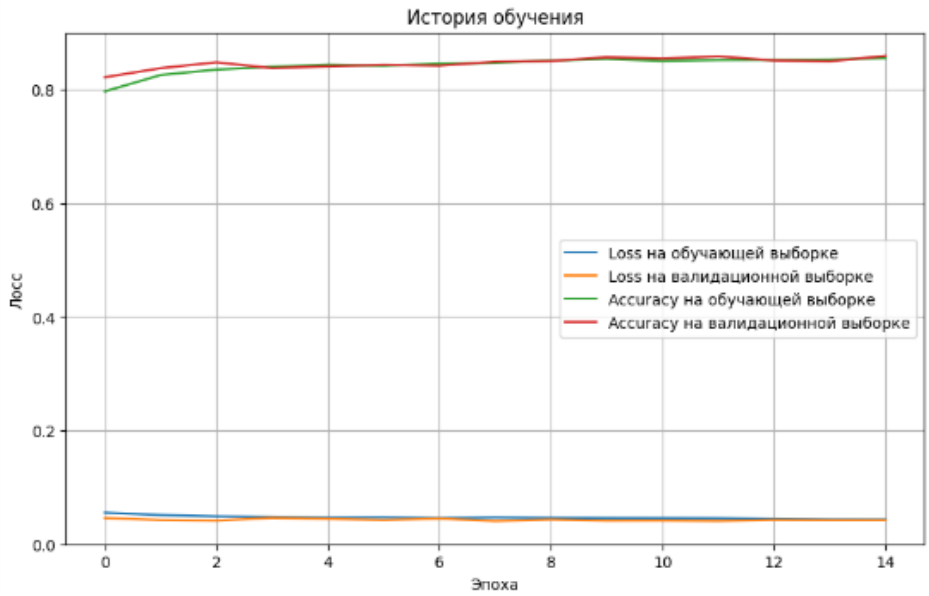

Рисунок 4.7 — История обучения MLP

4.4 ТЕСТИРОВАНИЕ МОДЕЛЕЙ

Тестирование моделей машинного обучения является важным этапом в процессе разработки, позволяющим оценить их производительность и обобщающую способность. В задаче бинарной классификации, где цель состоит в том, чтобы классифицировать объекты на два класса, тестирование моделей заключается в оценивании метрик.

Модель	**Точность**
MLP	0.853 ± 0.001
XGBoost	0.849 ± 0.002

Таблица 4.2 — Значение точности для обученных моделей на тестовой выборке

Так как в данной задаче одинаково важны точности принадлежности объекта к любому классу, основной метрикой для оценки служит accuracy и ROC-AUC. Значение точности для каждой модели представлено в таблице 4.2.

Также немаловажной оценкой производительности модели служит матрица ошибок. В данной задаче под значениями в матрице предполагается сумма весов при 4 различных вариантах.

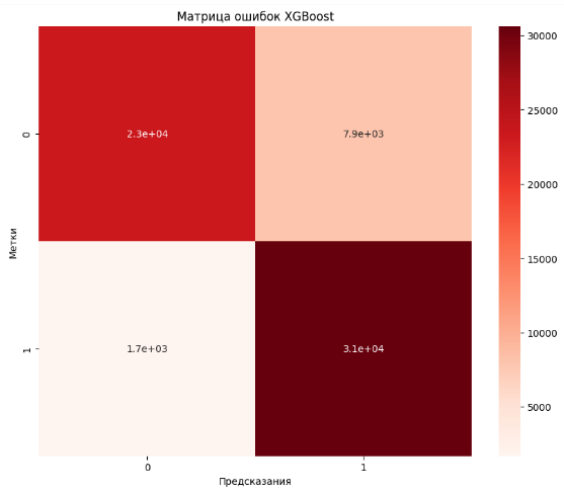

Рисунок 4.8 — Матрица ошибок для XGBoost

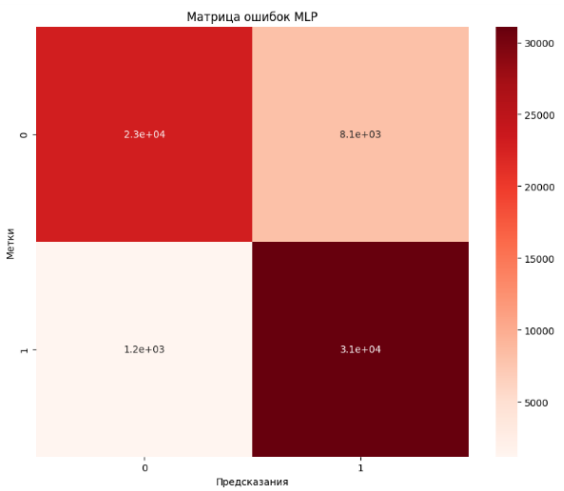

Рисунок 4.9 — Матрица ошибок для MLP

ROC-кривая позволяет сравнить и оценить результаты двух обученных моделей.

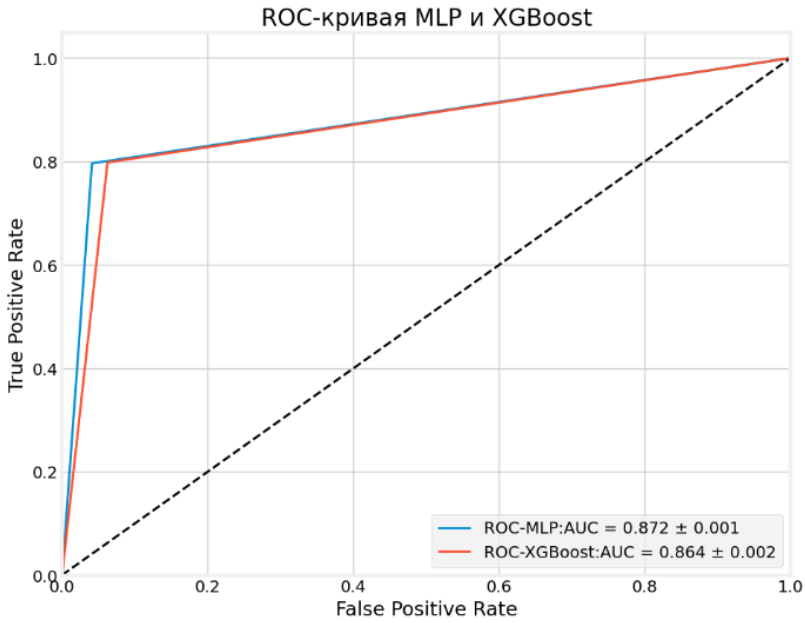

Рисунок 4.10 — ROC-кривые для моделей MLP, XGBoost для тестовых данных

4.4.1 ВЫВОД

По результатам тестирования можно сказать, что обе модели имеют обобщающие свойства и у них нет тенденции к переобучению. Несмотря на то, что в пределах погрешности модель MLP показывает себя лучше, их результаты сопоставимы.

При использовании модели XGBoost к реальным данным в качестве классификатора в распределении по инвариантной массе толстой струи в разрезе сигнала и фона можно судить о том, что данная модель справляется с идентификацией струй, образованных W-бозоном.

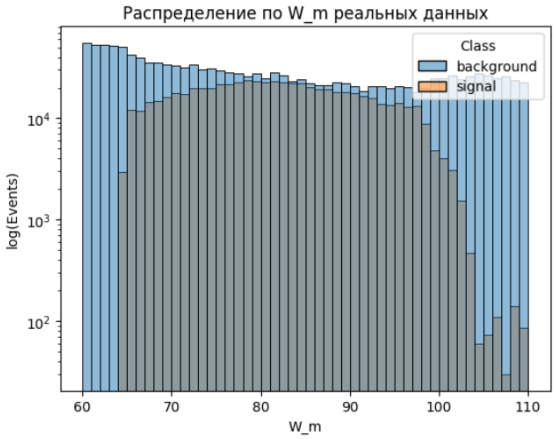

Рисунок 4.11 — Распределение по инвариантной массе струи для реальных данных в разрезе сигнала и фона

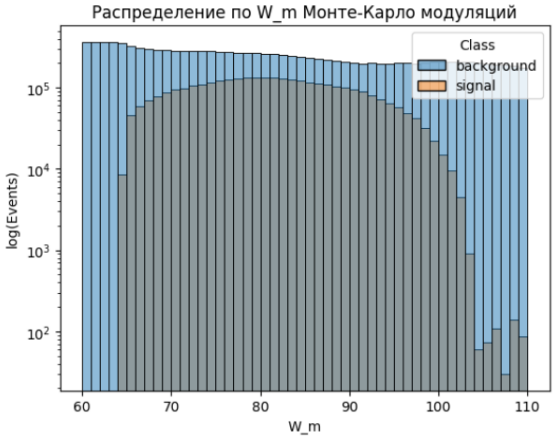

Рисунок 4.12 — Распределение по инвариантной массе струи для реальных данных в разрезе сигнала и фона

Однако в дальнейшем необходимо будет детальнее исследовать работу обеих моделей на реальных данных.

5 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.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Perkins D. H.* Introduction to high energy physics. — 1982. — ISBN 978-0-521-62196-0.
2. *Okun L. B.* Leptons and Quarks: Special Edition Commemorating the Discovery of the Higgs Boson. — Amsterdam, Netherlands : North-Holland, 1982. — ISBN 978-981-4603-14-0, 978-981-4603-00-3, 978-0-444-86924-1.
3. *Данных Ш. А.* Учебник по машинному обучению. — `https://education.yandex.ru/handbook/ml/article/about`.
4. *Chollet F.* Deep learning with python. — New York, NY : Manning Publications, 2017.
5. *Engineering S. U. S. of.* Свёрточные нейронные сети для визуального распознавания. — `https://www.reg.ru/blog/stenfordskij-kurs-lekciya-1-vvedenie/`.
6. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation. — 2008. — Т. 3, № 08. — S08003.
7. Identification of Hadronically-Decaying W Bosons and Top Quarks Using High-Level Features as Input to Boosted Decision Trees and Deep Neural Networks in ATLAS at $\sqrt{s} = 13$ TeV : тех. отч. / CERN. — Geneva, 2017. — All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atla PHYS-PUB-2017-004.
8. *Atkin R.* Review of jet reconstruction algorithms // J. Phys. Conf. Ser. / под ред. A. S. Cornell, B. Mellado. — 2015. — Т. 645, № 1. — С. 012008.
9. *Larkoski A. J.*, *Salam G. P.*, *Thaler J.* Energy correlation functions for jet substructure // Journal of High Energy Physics. — 2013. — Т. 2013, № 6. — ISSN 1029-8479.