Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский ядерный университет «МИФИ»

УДК 53.05, 53.07

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Классификация лептонных распадов W бозона методами машинного обучения в p-p столкновениях при $\sqrt{S} = 13$ TeB в эксперименте ATLAS

Научный руководитель _____ Д. Е. Пономаренко

Выполнил _____ Г. А. Толкачёв

Москва 2020

Оглавление

1 Введение

Исследования в физике элементарных частиц привели к созданию теории взаимодеиствия частиц на субъядерном уровне, которую принято называть Стандартной моделью. В Стандартной модели существует три поколения лептонов. Согласно лептонной универсальности предполагается, что три поколения лептонов во всех процессах должны вести себя одинаково. В рамках данного предположения отношения отношений сечений лептонного распада W бозона для любой пары лептонных каналов распада должны быть равно единице. В анализе данных с LEP имеются различия между теоретическим предсказанием Стандартной модели и экспериментальными измерениями[7]. А именно, существует указание на возможное отклонение в отношении отношений сечения двух процессов лептонного распада W бозона $(Br(W \to \tau\nu \to \mu\nu\nu)/Br(W \to \mu\nu))$ (рис. 1).

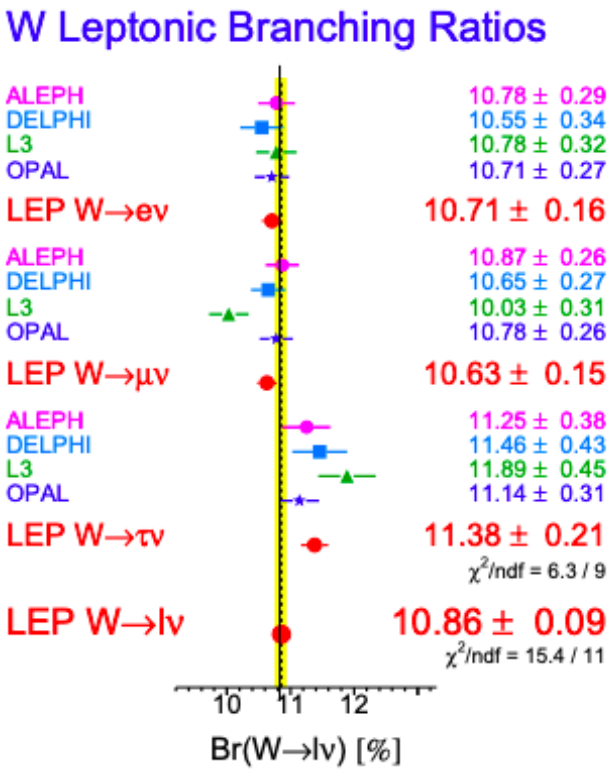


Рисунок 1 – Отношение сечения определенного канала распада к полному сечению распада W бозона

Анализ, в рамках которого проводится научно-исследовательская работа, является актуальным для поиска отклонения от предсказаний Стандартной

модели. Если в результате данного анализа полученный результат совпадет с теоретическим предсказанием, то это может свидетельствовать о подтверждении лептонной универсальности. В противном случае, если данный анализ подтвердит возможное отклонение в анализе данных с LEP, то это укажет на однозначное существование Новой физики. Данное, возможное отклонение от предсказаний Стандартной модели можно объяснить несколькими способами. Например, одно из возможных объяснений требует введения новой частицы, по-разному взаимодеиствующей с различными лептонами. Возможно, это новыи класс частиц, например, лептокварки или же тяжелыи аналог Z бозона.

Важным вкладом в данный анализ является работа над созданием и оптимизацией модели машинного обучения, с помощью которой можно провести классификацию данных. В результате классификации данных можно получить переменную, которая объединяет все кинематические переменные, используемые при классификации. На основе данной переменной удастся разделить очень кинематически схожие события. Таким образом, в дальнейшем можно будет провести более точное измерение отношения отношений сечений лептонных распадов W бозона.

Целью данной работы является нахождение оптимальной модели машинного обучения лес деревьев решений(BDT). А также исследование влияния точности измерения лептонной универсальности от использования отклика классификатора BDT в базовом анализе.

1.1 Стандартная модель

Исследования в физике элементарных частиц привели к созданию теории взаимодействия частиц на субъядерном уровне, которую принято называть Стандартной моделью[14]. Стандартная модель позволяет теоретически предсказать свойства различных процессов в физике элементарных частиц. В рамках Стандартной модели имеется 2 типа элементарных частиц: бозоны и фермионы. Фермионы имеют полуцелый спин. Сами фермионы делятся на две подгруппы: кварки и лептоны. Лептоны делятся на 2 типа: электрически заряженные частицы (e, μ, τ) нейтральные частицы — нейтрино (ν_e,

ν_μ, ν_τ). Кварки являются массивными частицами, имеющие электрический и цветовой заряд, всего их 6 (u, d, c, s, t, b). Все фермион по возрастанию массы делатся на три поколения. В стандартную модель входит 3 вида взаимодействия: электромагнитное, слабое и сильное. Сильное взаимодействие описывается квантовой хромодинамикой(КХД). Электромагнитное и слабое взаимодействия являются составными частями электро-слабого взаимодействия.

Лептонная универсальность

В рамках Стандартной модели существует три поколения лептонов. Три поколения упорядочены по массе заряженного лептона в диапазоне от 0.511 МэВ для e до 105 МэВ для μ, и 1.777 для τ [11]. Различие в массе приводят к совершенно разным временам жизни. От стабильного e до 2.2 мкс для μ, и 0.29 пс для τ. Лептоны участвуют в электромагнитных и слабых, но не сильных взаимодействиях, тогда как нейтрино участвуют только в слабом взаимодействии. Стандартная модель предполагает, что эти взаимодействия заряженных и нейтральных лептонов универсальны, т.е. одинаковы для трех поколений. Количество поколений лептонов пока не объяснено в рамках стандартной модели. Почти все наблюдаемые во Вселенной процессы выглядели бы точно так же, если бы существовало только одно поколение лептонов. Лептонную универсальность можно, например, проверить в лептонном распаде W бозона. Масса W достаточно большая, чтобы пренебречь массами лептонов, на которые он распадается. Поэтому между лептонными распадами W не должно быть разницы.

1.2 Машинное обучение

Роль машинного обучения в современном мире трудно переоценить, с помощью него можно выполнять различные задачи, как в физике высоких энергий, так и в других областях науки. В данной работе использовалась `TMVA` (Toolkit for Multivariate Data Analysis with `ROOT`) [13] — open-source библиотека алгоритмов машинного обучения, которая идёт в дополнение к пакету анализа больших данных `ROOT`. На основе TMVA в работе проводилось обучение модели BDT. После завершения обучения полученная модель BDT

используется для классификации данных. Полученный отклик модели BDT планируется использовать для проверки его валяния на финальное измерение лептонной универсальности в базовом анализе.

Классификатор — лес деревьев решений (BDT)

Метод классификации лес деревьев решений (BDT — Boosted Decision Trees) является одним из самый простых и популярных. На этапе обучения осуществляется отбор данных таким образом, чтобы на выходе получить максимальный прирост информации о данных, на каждом шаге выбирается тот признак, при разделении по которому прирост информации оказывается наибольшим. Дальше процедура повторяется рекурсивно.

Дерево решений — это метод представления решающих правил в иерархической структуре, состоящей из элементов двух типов — узлов и листьев. В узлах находятся решающие правила и производится проверка соответствия примеров этому правилу по какому-либо атрибуту обучающего множества (рис. 2). В простейшем случае, в результате проверки, множество примеров,

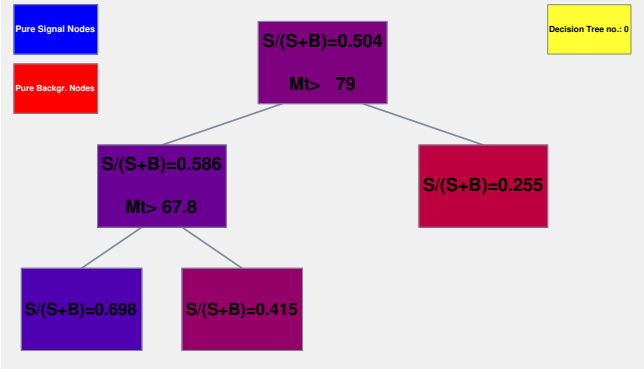


Рисунок 2 – Иллюстрация работы дерева решений

попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое — не удовлетворяющие. Затем к каждому подмножеству вновь применяется правило и процедура рекурсивно повторяется пока не будет достигнуто некоторое условие остановки алгоритма. В результате в последнем узле проверка и разбиение не производится и он объявляется листом.

Лес деревьев решений — это множество решающих деревьев. При построении каждого нового дерева учитывает «опыт» прошлых деревьев таким обра-

зом, что бы оно минимизировало ошибку всех предыдущих деревьев. При завершении обучения мы получаем дерево которое имеет минимальную ошибку при классификации. Такой подход называется градиентным бустингом. Важным критерием при построении леса деревьев решений является количество деревьев, а также глубина дерева.

ROC–integ

При классификации данных модель машинного обучения может давать как положительные ответы так и отрицательные. Все ответы классификатора могут быть записаны в таблицу, которая называется матрица ошибок(Confusion Matrix). Матрица ошибок состоит из 4 ячеек (таблица 1).

- Верно-положительные (TP), объекты, которые были классифицированы как положительные и деиствительно являются положительными (принадлежащими к данному классу).
- Верно-отрицательные (TN) объекты, которые были классифицированы как отрицательные и деиствительно отрицательные.
- Ложно-положительные (FP) объекты, которые были классифицированы как положительные, но фактически отрицательные.
- Ложно-отрицательныи (FN) объекты, которые были классифицированы как отрицательные, но фактически положительные

	$y = 1$	$y = 0$
$x = 1$	TP	FP
$x = 0$	FN	TN

Таблица 1 – Матрица ошибок

Где x - ответ алгоритма на объекте, y - истинная метка класса на этом объекте. Таким образом ошибки классификатора бывают видов FP или FN.

На основе матрицы ошибок и ее значений, рассчитываются различные метрики классификационной способности алгоритма. В данной работе для оценки качества бинарной классификации используют интеграл под ROC кривой(ROC - integ). Для построения ROC кривой для каждого классифицируемого объекта вычисляются две метрики оценки. Первой метрикой является специфичность(TNR или Background rejection), которая представляет отношение между верно классифицированных негативных экземпляров к числу всех негативных экземпляров. Вторая метрика является полнотой(TPR или Signal efficiency), которая представляет из себя пропорциию всех верноположительно предсказанных объектов к общему количеству деиствительно положительных. То есть, полнота показывает сколько образцов из всех положительных примеров были классифицированы правильно. Чем выше значение полноты, тем меньше положительных примеров пропущено в классификации.

$$TNR = \frac{TN}{TN + FP} \qquad TPR = \frac{TP}{TP + FN}$$

После вычисления данных метрик производится построение их на одном графике с осями Background rejection и Signal efficiency(рис. 3).

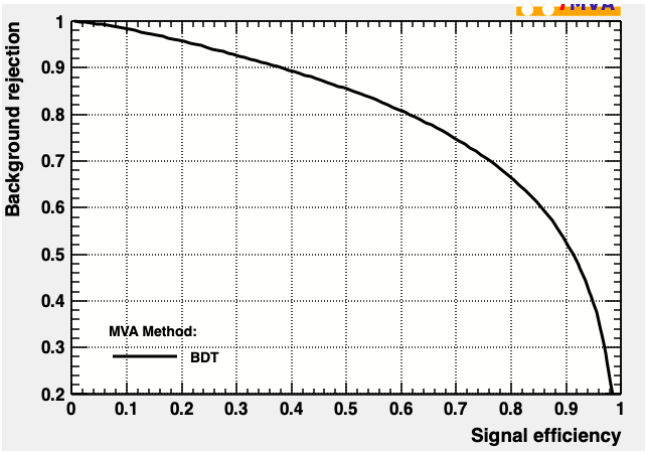


Рисунок 3 – График ROC. Signal efficiency - TPR. Background rejection - TNR

После чего вычисляется интеграл под ROC кривой. Чем больше интеграл под ROC кривои, тем лучше модель классифицирует события.

2 Детектор ATLAS

ATLAS (от англ. A Toroidal LHC ApparatuS) — один из четырёх основных экспериментов на коллайдере LHC в Европейской Организации Ядерных исследований CERN в городе Женева (Швейцария) (рис. 4). Эксперимент

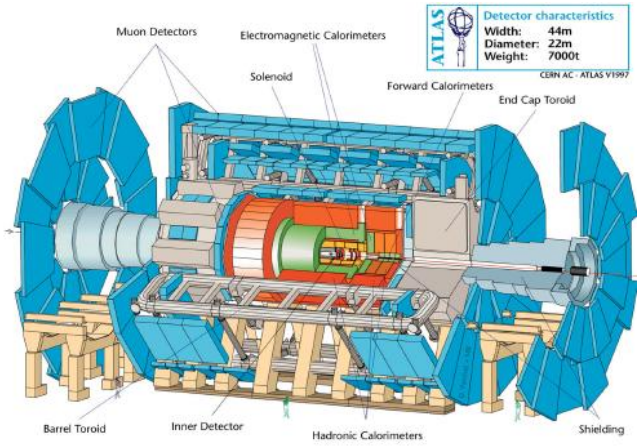


Рисунок 4 – Детектор ATLAS

проводится на одноимённом детекторе, предназначенном для исследования протон-протонных столкновений. Детектор состоит из нескольких частей. Для восстановления треков и импульсов заряженных частиц используется внутренний детектор, окруженный сверхпроводящим магнитом, создающим магнитное поле 2 Тл. С помощью системы калориметров происходит измерение энергии частиц. На периферии детектора находится мюонный спектрометр, который необходим для измерения импульса мюонов. Для отбора событий используется система триггеров.

3 Программное обеспечение

Один из инструментов для проведения анализа в физике высоких энергий является языки программирования. Однако, написание отдельных скриптов на основе языков программирования для каждого анализа является достаточно расточительным. Хорошим решением для проведения анализа данных в физике высоких энергий является создание программного обеспечения, которое может быть использовано не только для конкретного анализа. При проведении анализа в данной работе было использовано программное обеспечение $xTauReadar$[4].

3.1 $xTauReadar$

Для проведения исследования кинематических переменных, а также для проведения тренировки модели машинного обучения и для оценки данных с помощью полученной модели был использован $xTauReader$ фреймворк. Данное программное обеспечение было разработано на основе библиотек `pyROOT` и пакета `HAPPy`[1] специально для проверки лептоннои универсальности в данных с эксперимента ATLAS. В рамках данного программного обеспечения был проведен ряд модернизаций существующего в $xTauReadar$ модуля. Модернизировался модуль, предназначенный для тренировки модели машинного обучения, а также для проведения классификации данных. Главной целью данных модернизаций является ускорение обработки данных методами мультипоточности на этапе обучения модели и оценки данных. Для повышения производительности был использована библиотека `ROOT::RDataFrame`.

3.2 ROOT::RDataFrame

`RDataFrame` является библиотекой пакета `ROOT`, которая предоставляет интерфейс высокого уровня для анализа данных, хранящихся в различных источниках, включая файлы root, которые в основном используются для хранения данных с экспериментов CERN. Он выполняет низкоуровневую оптимизацию и кэширование выполняемых вычислений и добавляет неявные возможности многопоточного распараллеливания. Благодаря использованию

многопоточного режима в `RDataFrame` можно значительно увеличить производительность, а также увеличить объем данных, обрабатываемых одновременно в единицу времени. `RDataFrame` хорошо себя зарекомендовал в анализе данных с эксперимента TOTEM[5].

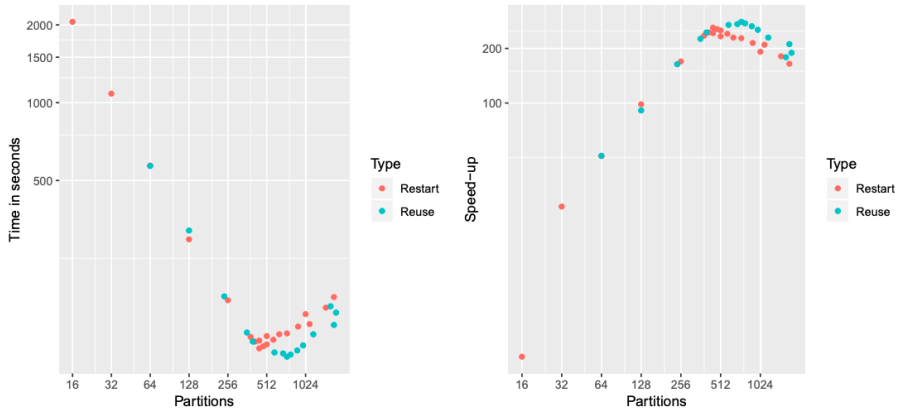


Рисунок 5 – Графики зависимости производительности от используемых ядер

На данном графике(рис. 5) отлично видно повышение производительности при увеличении количества ядер, использованных в анализе данных с эксперимента TOTEM.

3.3 Модернизация xTauReader

В процессе выполнения данной работы было произведено улучшение имеющегося модуля *xTatuReader*, а именно модуля отвечающего за работу с пакетом ROOT::TMVA и предназначенного для выполнения обучения модели и классификации данных. В первую очередь был произведен переход на многопроцессорную обработку на этапе классификации данных, для этого и использовалась библиотека RDataFrame. Однако, при тестировании новой реализации функций класса, была выявлена проблема. Данная проблема связанна с тем, что после обработки данных с помощью RDataFrame в многопоточном режиме результирующий файл содержит порядок событий, который отличается от первоначального файла. Это происходит из-за того, что при использовании `RDataFrame` в многопоточном режиме один набор данных обрабатывается параллельно на нескольких потоках и очередность выполнения на этих потоках ничем не контролируется, из-за чего может нарушаться

последовательность событий в финальном файле. Эта проблема является весомой из за того, что после проведения классификации набора данных, полученный отклик записывается на носитель. Далее полученный файл объединяется методом `ROOT::TTree::AddFriend` с исходными данными, на которых была осуществлена классификация. После объединения файлов производится отбор событий. Если порядок событий в файле с откликом модели отличается от порядка в изначальном наборе данных, то при проведении отбора событий каждое событие в изначальном файле будет не соответствовать событию в файле с откликом модели. Таким образом отбор событий будет произведен неверно. Данная проблема была решена при помощи библиотеки `multiprocessing`, благодаря которой можно так же добиться оптимизации процесса, при этом избежать перемешивания событий. Происходит это из за того, что с помощью библиотеки `multiprocessing` можно обрабатывать несколько наборов данных, обработка каждого из которых будет производится в отдельном потоке.

Помимо проблемы с мультипоточностью была выявлена проблема с переполнением памяти, которая связана в первую очередь с объёмом обрабатываемых данных. При обучение модели с использованием `TMVA` необходимо загрузить данные с помощью метода `TMVA::Factory::AddTree`, при этом использованные данные загружаются в оперативную память вычислительной машины. Из за того, что объем использованных данных иногда может превышать объем оперативной памяти, программа может завершить свою работу из-за ошибок, возникающих по причине нехватки памяти. Для снижения объема данных перед каждой тренировкой модели или классификацией данных было принято решение делать перезаписать изначальных файлов в временные файлы, которые бы имели только необходимые переменные. Для этого была написана функция, которая выполняет поиск в формулах переменных, используемых на этапе обучения или классификации. С помощью чего можно произвести запись файлов, которые будут содержать только переменные, необходимые для проведения обучения модели или классификации.

4 Использованные данные

Во время набора данных на эксперименте ATLAS рождается огромное число фоновых событий, для того чтобы понять какой вклад они вносят применяют статистический анализ, важным компонентом которого являются данные, сгенерированные методом Монте-Карло[9]. В работе использовались данные, полученные с использованием $HistMaker$[2].

4.1 Экспериментальные данные

В работе использовались экспериментальные данные, набранные на детекторе ATLAS в 2017 и 2018 году во время режима набора данных с низкой светимостью 340 пБ$^{-1}$. При столкновении протон-протонных пучков с суммарной энергией 13 ТэВ.

4.2 Монте-Карло моделирование

Смоделированные данные, используемые в работе, были получены методом Монте-Карло с помощью генераторов Pythia[12] и Sherpa[8] и прошли всю цепочку реконструкций, на условии реальных протон-протонных столкновений эксперимента ATLAS сессия 2 (RUN 2). Каждому каналу соответствует свой уникальный номер DSID. Список Монте-Карло данных, использованных в работе, приведен в таблице 2

Для сравнения Монте–Карло и реальных данных выполняется нормировка на светимость. Для более точного согласия с распределениями из данных используются коррекционные коэффициенты, которые учитывают неточности в моделирование Монте—Карло и геометрию детектора. Коэффициенты коррекции, использованные в работе предоставляются Combined Performance (CP) Groups эксперимента ATLAS [6].

Sample	DSID	Generator	xs [pb]
$W^+ \to \mu\nu$	361101	PowhegPythia8EvtGen	11500.9
$W^- \to \mu\nu$	361104	PowhegPythia8EvtGen	8579.31
$W^+ \to \tau\nu$	361102	PowhegPythia8EvtGen	11500.9
$W^- \to \tau\nu$	361105	PowhegPythia8EvtGen	8579.31
$Z \to \tau\tau$	361108	PowhegPythia8EvtGen	1950.63
$Z \to \mu\mu$	361107	PowhegPythia8EvtGen	1950.63
Top	410013	PhPy8EG_P2012	35.8455
Top	410014	PhPy8EG_P2012	35.8244
Top	410470	PhPy8EG	729.77
Top	410642	PhPy8EG	36.993
Top	410643	PhPy8EG	22.174
Top	410644	PowhegPythia8EvtGen	2.06146
Top	410645	PowhegPythia8EvtGen	1.28867
Diboson	363356	Sherpa_221_PDF30	2.20355
Diboson	363358	Sherpa_221_PDF30	3.4328
Diboson	363359	Sherpa_221_PDF30	24.708
Diboson	363360	Sherpa_221_PDF30	24.724
Diboson	363489	Sherpa_221_PDF30	11.42
Diboson	364250	Sherpa_221_PDF30	1.2523
Diboson	364253	Sherpa_221_PDF30	4.579
Diboson	364254	Sherpa_221_PDF30	12.501
Diboson	364255	Sherpa_221_PDF30	3.2344

Таблица 2 – Список каналов, использованных в генераторе Монте-Карло при моделировании данных.

5 Предварительный отбор

В данной работе используется два контрольных региона. Однин является сигнальным регионом, а второй является Z регионом. Каждый регион характеризуется отдельным набором ограничений на кинематические переменные и сигнатуру событий.

5.1 Z регион

В данном анализе Z регион используется как контрольный регион. В нем экспериментальные данные очень хорошо согласуются с Монте-Карло данными, а также содержится малое количество КХД фона. С помощью Z региона проводят валидацию сигнального региона и используют для уменьшения систематической погрешности.

Отбор событий Z региона

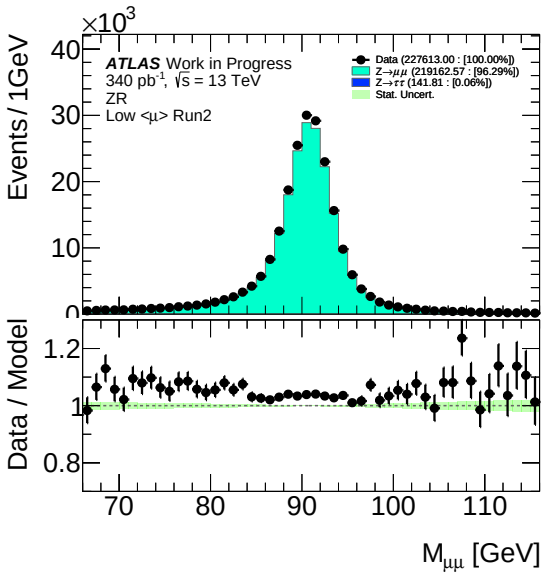
Для получения Z региона был проведен отбор, который включает в себя ограничение на поперечный импульс P_T больше 20 ГэВ для двух лептонов. Отбор на инвариантную массу от 66 до 116 ГэВ. А также были применены критерии на изоляцию лидирующего лепота и идентификацию.

P_T	M_{ll}	ID	Isolated
$>$ 20 ГэВ	66 ГэВ $<$ и $<$ 116 ГэВ	tight	medium

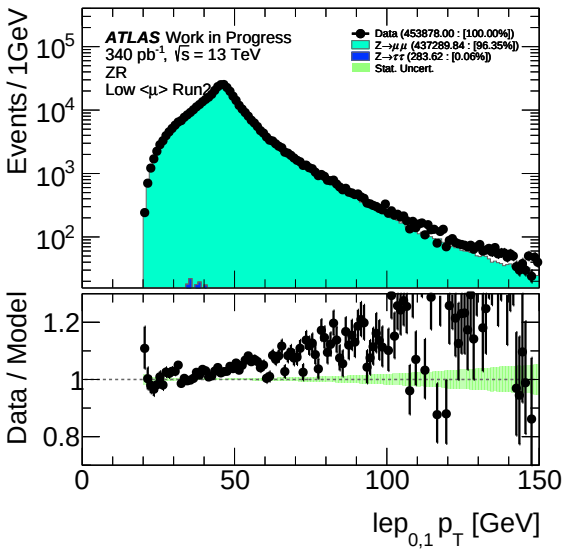
Таблица 3 – Отбор событий Z региона

Кинематические распределения Z региона

Для Z региона было выполнено сравнение кинематических распределений для реальный и Монте—Карло данных, а также показано их отношение. В Z регионе содержится мало КХД фона и Монте—Карло хорошо согласуются с экспериментальными данными, однако, на представленном отношении



(а) Инвариантная масса M_{ll}



(б) Поперечный импульс P_T каждого лептона из пары

Рисунок 6 – Сравнение реальных и Монте-Карло данных Z региона для переменных: $M_{\|}$ и поперечного импулса каждого лептона P_T из пары

имеются отклонения Монте—Карло от экспериментальных данных в жесткой части поперечного импульса каждого лептона из пары $lep_{0,1}P_T$(рис. 6б).

5.2 Сигнальный регион

Сигнальный регион является основным в данной работе. Он содержит события с распадами W бозона на τ лептон. На данных сигнального региона производится обучение модели машинного обучения. А также с использованием данных сигнального региона получают отклик модели, который в дальнейшем планируется использовать для финального измерения лептонной универсальности в базовом анализе. Для отбора сигнального региона отбираются событии, которые имеют схожую сигнатуру с распадом W бозона в τ лептон.

Отбор событий сигнального региона

Для получения сигнального региона был проведен отбор, который включает в себя ограничение на поперечный импульс P_T больше 20 ГэВ. Отбор на поперечную энергию E_T^{miss} больше 20 ГэВ, поперечную массу M_T больше 40

ГэВ, определяемую как

$$m_T = \sqrt{2P_T(l)P_T(\nu)(1 - \cos(\phi(l) - \phi(\nu)))}, \quad (1)$$

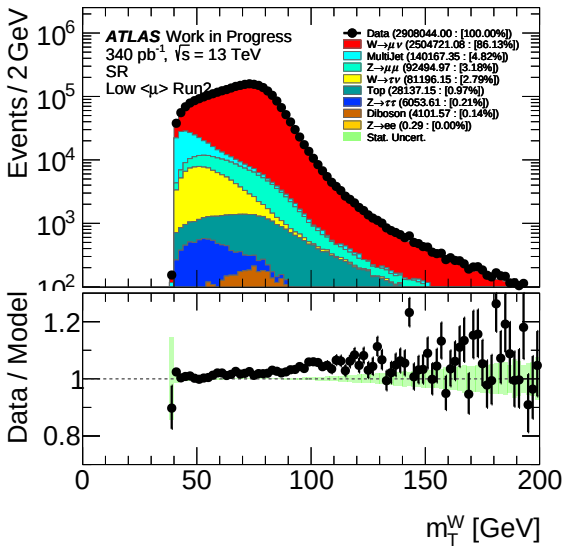
Данный отбор проводится для отсечения большого количества КХД фона. Также использовалось ограничение на количество лептонов в событии равное 1. Были применены критерии идентификацию мюона - Tight, а также на изолированность мюона - Medium(таблица 4). Для улучшения выделения настоящих лептонов от других, например, не интересующих нас частиц или струй, к лептонам применяется отбор на изолированность. Трековая изоляция рассчитывается как сумма поперечных импульсов всех треков в конусе размера $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$, кроме трека самого лептона, и делится на импульс лептона. В работе используется трековая изоляция в конусе $\Delta R = 0.2$, равная ptvarcone20/P_T<0.1.

P_T	E_T^{miss}	m_T	N lep	ptvarcone20/P_T	ID	Isolated
$>$ 20 ГэВ	$>$ 20 ГэВ	$>$ 40 ГэВ	1	$<$ 0.1	tight	medium

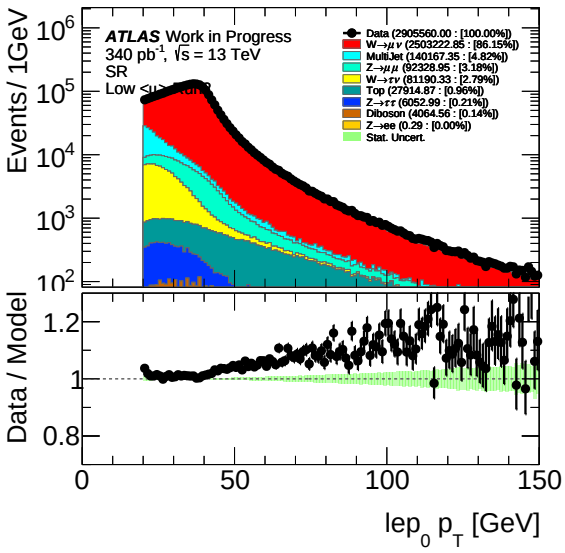
Таблица 4 – Отбор событий сигналного региона

Кинематические распределения сигнального региона

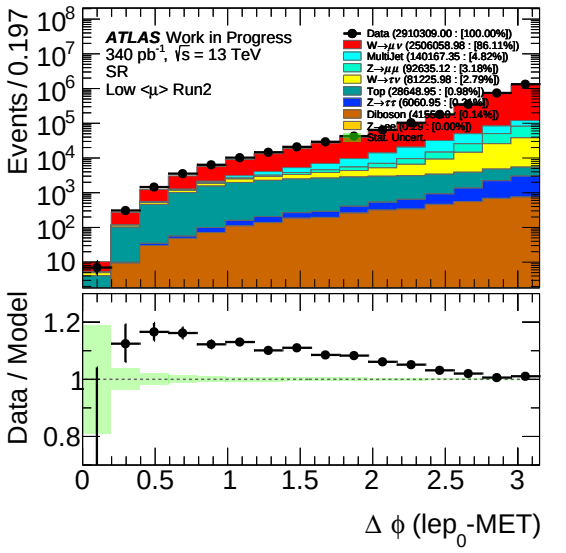
Было выполнено сравнение кинематических распределений для реальных и Монте-Карло данных (рис. 7 и 8), а также показано отношение реальных данных к данным Монте-Карло. По данному распределению видно, что по-



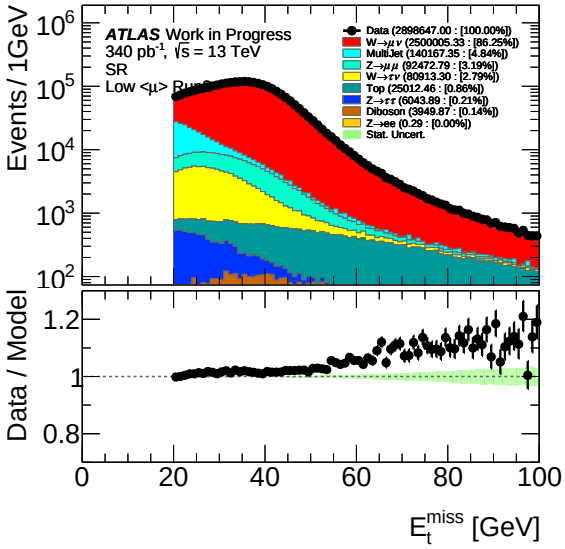
(а) Поперечная масса m_T



(б) Поперечный импульс P_T



(в) Разность углов ϕ потерянной энергии и лептона $d\phi(l - E_{\text{miss}})$

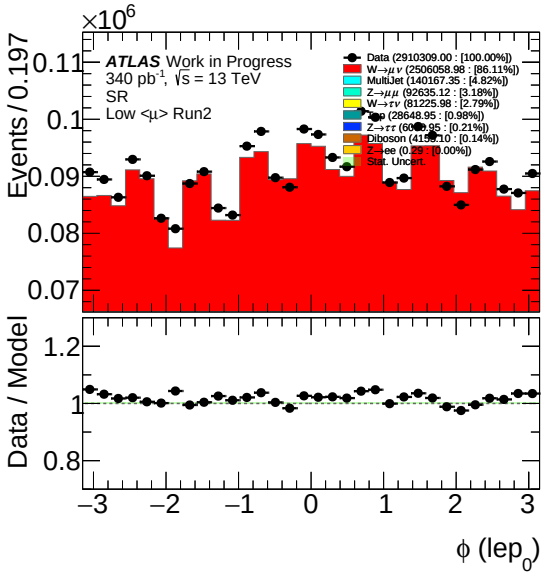


(г) Потерянная поперечная энергия E_{miss}

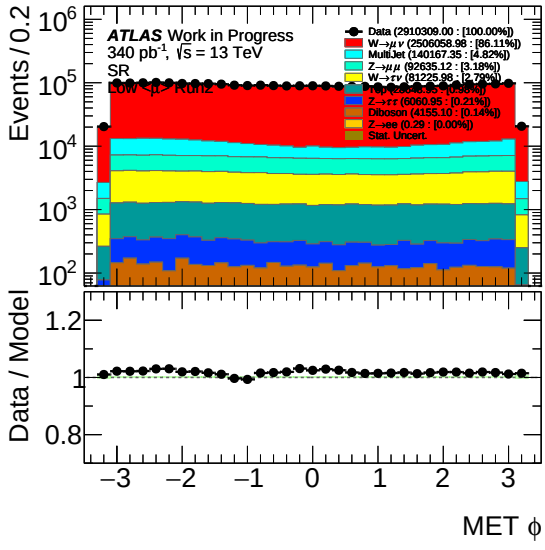
Рисунок 7 – Сравнение реальных и Монте-Карло данных для переменных: M_T, P_T , $d\phi(l - E_{\text{miss}})$, E_{miss}

мимо сигнального процесса большой вклад вносят фоновые события. Наибольший вклад вносит процесс распада W бозона в лептон второго поколе-

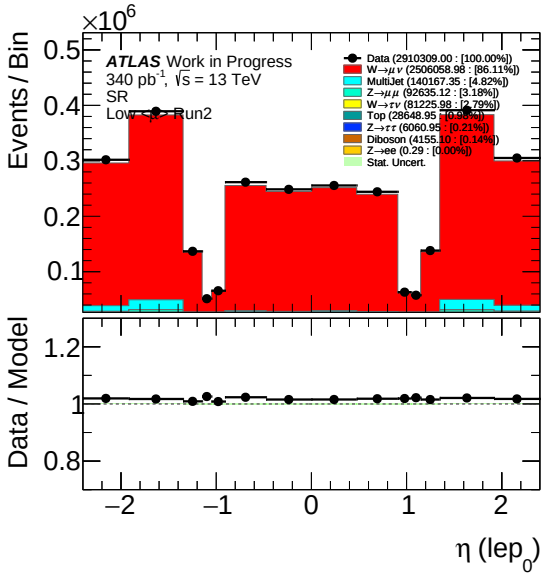
ния, значительный вклад вносит также КХД фон и распады Z бозона на два лептона. Участие распада Z бозона в фоновым процессе происходит из-за того, что иногда один из лептон не удается зарегистрировать. Поэтому данный распад имеет похожую сигнатуру с распадом W бозона в лептон. На представленном отношении отчетливо видно отклонение смоделирован-



(а) Угол ϕ лептона



(б) Угол ϕ потерянной энергии



(в) Псевдобыстрота η

Рисунок 8 – Сравнение реальных и Монте-Карло данных для переменных: lep ϕ, MET ϕ, lep η

ных Монте-Карло данных от реальных данных в жесткой части поперечного импульса P_T, потерянной энергии E_T^{miss}, поперечной массы M_T. Данное отклонение имеется также в Z регионе. Поэтому можно сделать вывод о том,

что данное отклонение связано с плохим моделированием Монте-Карло в генераторе PowhegPythia.

Оценка КХД фона

Используемые в данной работе генераторы Монте—Карло не могут надежно смоделировать КХД фон. Поэтому необходимо произвести оценку КХД фона с помощью метода оценки фона из данных(data-driven method)[10]. КХД фон был посчитал анализ группой[3] и предоставлен в мою работу для использования в исследовании отклика BDT моделей.

5.3 Псевдо–W регион

Помимо физических регионов в работе также используется псевдо-W регион, который является искусственно созданным регионом и используется для валидации модели машинного обучения. Для создания псевдо-W региона для каждого события с вероятностью 50 процентов был выбран один из двух лептонов Z региона. После этого выбранный лептон был записан как лидирующий, а второй лептон был добавлен к первоначальной потерянной энергии формула 2.

$$E^{miss}_{T,pseudoW} = E^{miss}_{T} rand(lep), \tag{2}$$

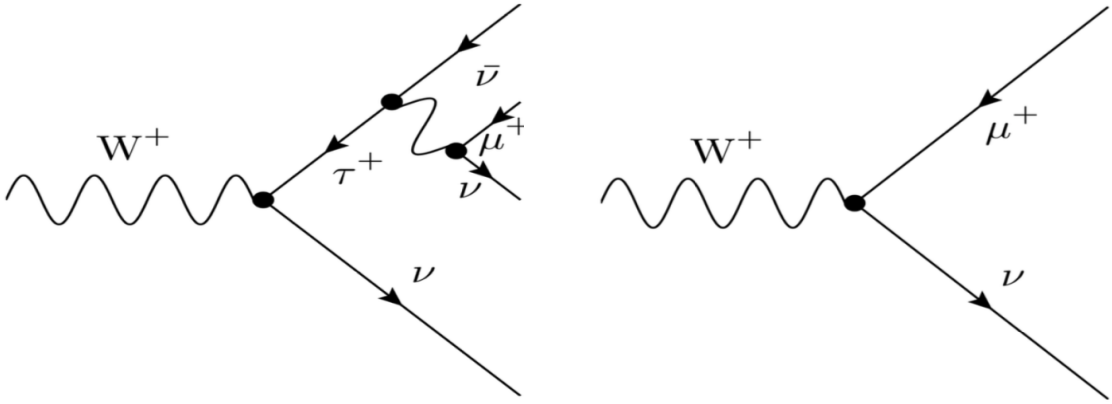
Псевдо-W регион имеет такой же отбор по кинематическим распределениям как и Z регион. Благодаря такой перезаписи кинематических переменных Z региона и отбору по кинематическим переменным мы имитируем сигнатуру сигнального региона, сохраняя при этом хорошее соответствие Монте—Карло и экспериментальных данных. Благодаря чему псевдо-W регион может быть использован для валидации отклика модели машинного обучения.

6 Обучение модели

В процессе обучения модели BDT был проведен предварительный отбор. Для повышения статистики сигнальных событий был использован только отбор на количество лептонов, равное одному и поперечный импульс P_T леп-

тона, больше 20, а также на изоляцию (Medium) и идентификацию(tight) лептона.

В качестве сигнала рассматриваются распады W в τ с последующим распадом в лептон второго поколения (рис. 9а), а в качестве фона рассматриваются прямые распады W в лептон второго поколения (рис. 9б), распады t кварка, дибозонные процессы, распады Z бозона на лептоны. Полученная выборка для сигнального региона содержит 182806 событий, а в свою очередь полученная выборка для фоновых событий содержит 732000 событий. Соотношение тренировочных и тестовых данных было выбрано 70 на 30 про-



(а) Распад W через τ (б) Прямой распад W

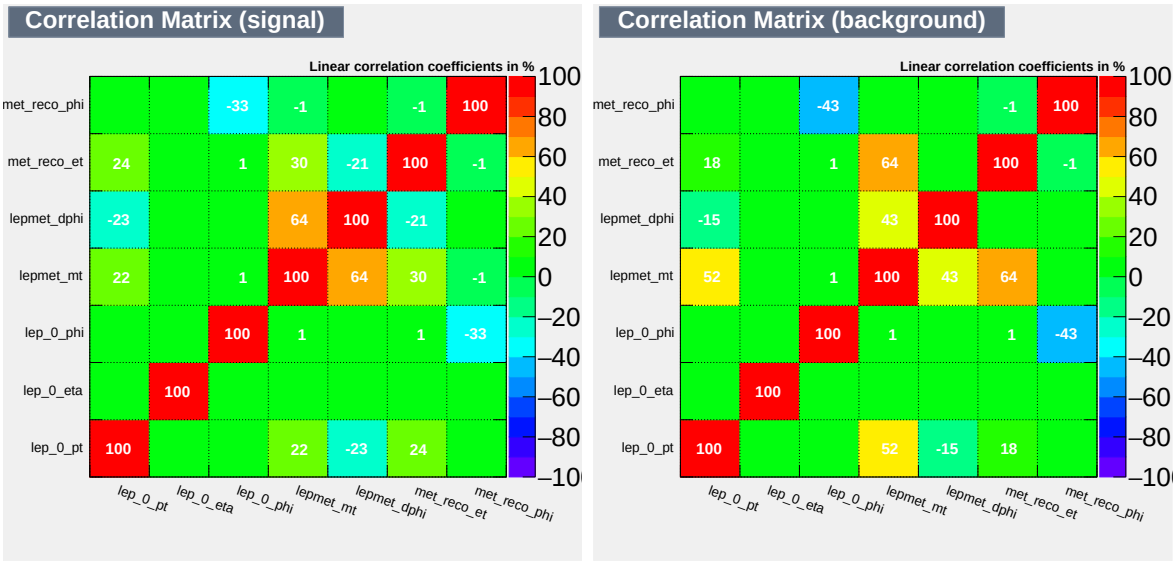
Рисунок 9 – Фейнмановские диаграммы

центов. В обучении модели BDT использовались следующие переменные:

- Поперечная масса m_T.
- Поперечный импульс P_T
- Псевдобыстрота лептона η
- Угол потерянной энергии MET ϕ
- Угол лептона ϕ_{lep}
- Поперечная потерянная энергия E_{miss}.
- Разность углов ϕ для лептона и потерянной энергии $d\phi(lep - MET)$.

Основываясь на результатах работы прошлого семестра, в которой проводилось исследование зависимости значения интеграла под ROC кривой от параметров модели BDT, были выбраны наилучшие параметры модели BDT. Этими параметрами являются количество деревьев - 200 и глубина деревьев -5.

Были построены матрицы корреляции для всех переменных, используемых в обучении (рис.10). Корреляция поперечной массы M_T с поперечным



(а) Матрица корреляции для сигнала (б) Матрица корреляции для фона

Рисунок 10 – Матрицы корреляции для сигнала и фона

импульсом P_T для сигнальных событий составляет 22 процента, а для фоновых 52 процента. В свою очередь, корреляция поперечной массы M_T с потерянной энергии E_T^{miss} для сигнальных событий составляет 30 процентов, а для фоновых 64 процентов. Так же имеется корреляция между M_T и разностью углов лептона и потерянной энергии $d\phi(lep - MET)$. Для сигнала корреляция составляет 64 процента, а для фона 43 процента. Однако, большая корреляция между поперечной массой M_T и переменными поперечного импульса P_T, потерянной энергии E_T^{miss} и разности углов $d\phi(lep - MET)$ является вполне ожидаемой(формула 1) и не является критичной.

Для определения вклада, который вносит каждая переменная при классификации, была посчитана значимость для каждой переменной, использованной при обучении модели BDT (таблица 5). Наибольший вклад в классификацию вносит поперечная масса M_T. Но, это является ожидаемым ре-

Переменная	P_T	M_T	$\Delta\phi$	E_T^{miss}	η	MET ϕ	ϕ_{lep}
Значимость	1.8e-01	2.6e-01	4.4e-02	1.2e-01	2.3e-04	1.2e-04	1.3e-04

Таблица 5 – Значимость переменных

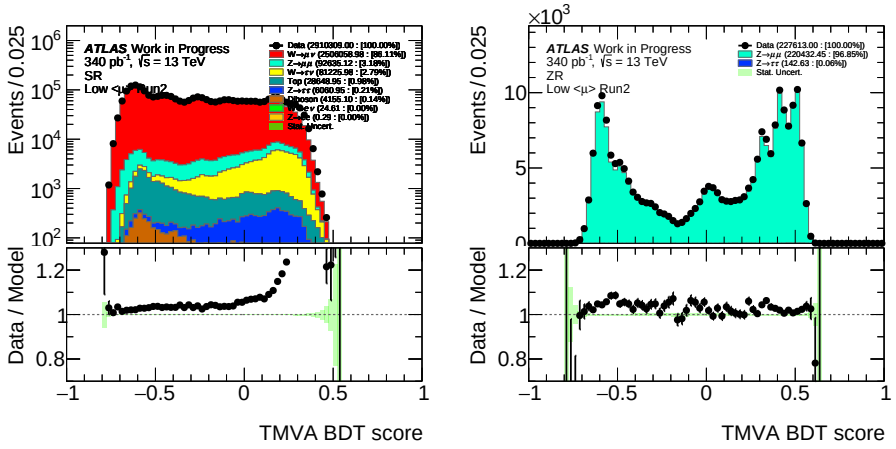
зультатом для поперечной массы M_T из-за ее зависимости от поперечного импульса P_T и потерянной поперечной энергии E_T^{miss}, которые сами вносят значительный вклад при классификации. Наименьший вклад вносят псевдобыстрота лептона lep η и азимутальный угол для потерянной энергии MET ϕ и лептона ϕ_{lep}.

Значение интеграла под ROC кривой после тренировки составляет 0.795.

7 Результат классификации событий

Готовая модель BDT использовалась для классификации данных сигнального региона. На первом этапе проводилась классификация проводилась без учета ХКД фона из-за того, что оценка КХД фона не была предоставлена в мой анализ. На представленном отношении отчетливо видно отклонение Монте-Карло от реальных данных(рис. 11а). Отклонения в области от -0.9 до 0 связаны с плохим моделированием данных Монте-Карло, например в распределении поперечного импульса P_T и потерянной поперечной энергии E_T^{miss} лептона. Имеются также отклонения в области от 0 до 0.5. Предполагается, что они были вызваны неучтенным КХД фоном. Для подтверждения данной гипотезы была проведена классификация данных псевдо-W региона.

По данному распределению (рис. 11б) можно сделать вывод, что классификация Монте-Карло и экспериментальных данных хорошо согласуется между собой, а это значит, что причиной отклонения в классификации данных сигнального региона в области 0 до 0.5 с большой вероятностью является неучтенный КХД фон. На втором этапе, после как оценка для КХД была посчитана и предоставлена в данный анализ, осуществлялась классификация с учетом КХД фона. На рисунке 12а представлен отклик модели на данные сигнального региона вместе с КХД фоном. Имеются отклонения Монте—Карло от реальный данных в области от 0.3 до 0.5. Причиной откло-

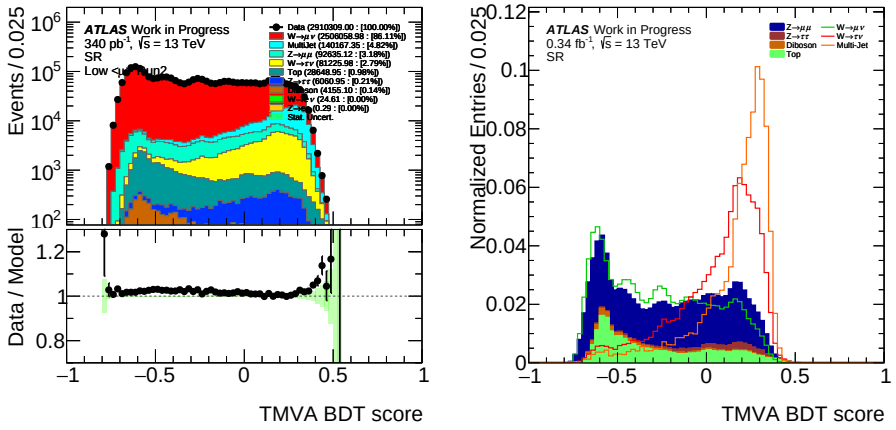


(а) Классификация данных сигнального региона

(б) Классификация данных псевдо-W региона

Рисунок 11 – Классификация данных

нений может являться как плохое моделирование Монте—Карло, так и плохая оценка КХД фона. В дальнейшем для выяснения данной причины планируется исследование кинематических переменных, используемых при обучении модели и классификации данных с добавлением еще одного ограничения. А именно ограничение на переменную скоринга модели больше 0.3.



(а) Классификация данных сигнального региона с КХД фоном

(б) Сравнение отклика модели BDT сигнального региона, после нормировки на интеграл.

Рисунок 12 – Отклик модели BDT

Также были построены нормированные на интеграл распределения отклика для каждого канала Монте—Карло данных (рис 12б). На данных распределениях видно явное отличие формы распределения сигнала от фона.

Форма распределений отклика для сигнала и КХД является схожей. Это происходит из-за того, что эти процессы имеют схожую сигнатуру, например, в распределении поперечного импульса P_T(рис. 7б) эти процессы находятся в мягкой части. То же самое можно сказать и о схожести всех фоновых событий, кроме КХД фона. На основе данного исследования можно сделать вывод о том, что модель BDT хорошо справляется с задачей классификации. Из-за того, что сигнальный и фоновые процессы хорошо различаются между собой, использование распределения отклика вместо поперечной массы m_T может улучшить точность измерения лептонной универсальности в базовом анализе.

8 Заключение

В процессе выполнения данной научной работы были построены кинематические распределения для Монте-Карло данных и реальных данных 2017 и 2018 года с энергией 13 ТэВ и светимостью 340 $пБ^{-1}$ эксперимента ATLAS. Показано расхождение в жесткой части поперечного импульса P_T, при больших значениях потерянной поперечной энергии E_T^{miss} в Z регоине и сигнальном регионе. Также имеются расхождения в поперечной массе m_T^W в сигнальном регионе. Данные расхождения связаны с плохим моделированием Монте-Карло данных в генераторе PowhegPythia.

Выполнена тренировка модели BDT, параметры которой были взяты из работы прошлого семестра. Далее был получен отклик классификатора событий для экспериментальных данных и данных Монте-Карло сигнального региона без оценки ХКД фона . Исследованы наблюдаемые расхождения между между Монте-Карло и экспериментальными данными с помощью псевдо-W региона. Сделано предположение о том, что отклоениня в области от 0 до 0.5 вызваны неучтенным КХД фоном. После предоставления КХД фона в данный анализ была проведена классификация событий с учетом КХД фона. Сделан вывод о оставшемся расхождении в области от 0.3 до 0.5. Причиной отклонения может являться как плохое моделирование Монте-Карло, так и плохая оценка КХД фона. Для выяснения причины отклонения планируется исследование кинематических переменных, используемых при обучении модели и классификации данных с добавлением ограничения на переменную оценки.

Помимо прочего, в процессе выполнения данной работы были изучены пакеты для проведения анализа данных $xTauReader$ и $HAPPy$. А также был выполнен ряд работ, которые связаны с модернизацией $xTauReader$ фреймворка. За счет данных улучшений удалось значительно повысить производительность на этапе классификации данных. Была также устранена ошибка, связанная с переполнением оперативной памяти, при использовании библиотечки `ROOT::TMVA`.

В качестве следующего шага работы планируется использование полученной переменной отклика модели BDT в измерении лептонной универсально-

сти и проверки ее влияния на точность измерения.

Список использованных источников

1. ATLAS HAPPy software documentation. — URL: https://gitlab.cern.ch/Wlep_BR/HAPPy/-/blob/master/README.md.
2. ATLAS HistMaker software documentation. — URL: https://gitlab.cern.ch/atlas-wbr-lowmu/HistMaker/-/blob/master/README.md.
3. ATLAS WBR analysis with low mu. — URL: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/Vtaus13TeV.
4. ATLAS xTauReader software documentation. — URL: https://gitlab.cern.ch/Wlep_BR/xTauReader/blob/master/doc/README.md.
5. *Blaszkiewicz Milosz (AGH-UST C.* Interactive data analysis of data from high energy physics experiments using Apache Spark : тех. отч. / CERN LHC ; TOTEM. — Geneva, 2019. — CERN-THESIS-2019—004.
6. Combined Performance (CP) Groups. — URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AtlasPhysics.
7. Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP / S. Schael [et al.] // Phys. Rept. — 2013. — Vol. 532. — P. 119–244.
8. Event Generation with Sherpa 2.2 / E. Bothmann [et al.] // SciPost Phys. — 2019. — Vol. 7, no. 3. — P. 034.
9. Monte Carlo event generators for high energy particle physics event simulation / S. Alioli [et al.]. — 2019. — Feb.
10. Multi-jet background in low-pile-up runs taken in 2017 and 2018 : тех. отч. / T. Xu [и др.] ; CERN. — Geneva, 07.2019. — ATL-COM-PHYS-2019—076.
11. Precision measurement of the mass of the τ lepton / M. Ablikim [et al.] // Phys. Rev. — 2014. — Vol. D90, no. 1. — P. 012001.

12. *Sjostrand T.*, *Mrenna S.*, *Skands P. Z.* A Brief Introduction to PYTHIA 8.1 // Comput. Phys. Commun. — 2008. — Vol. 178. — P. 852–867.

13. TMVA - Toolkit for Multivariate Data Analysis / A. Hoecker [et al.]. — 2007.

14. *M. E. B.* Стандартная модель и её расширения. — НМ. : Физматлит, 2007. — 584.