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

Студент: Ван Алина Маошэновна, М23-112

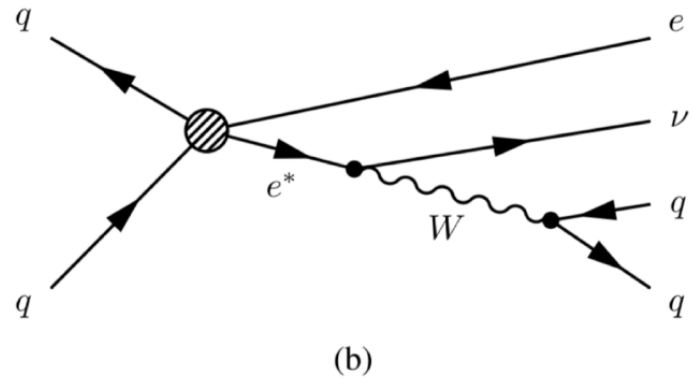
Научный руководитель: Мягков Алексей Григорьевич, к.ф.-м.н.

2025

Мотивация

Проблемы Стандартной модели:

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



Пример: Поиск возбужденного лептона с последующим распадом через калибровочный бозон

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

Цель и задачи

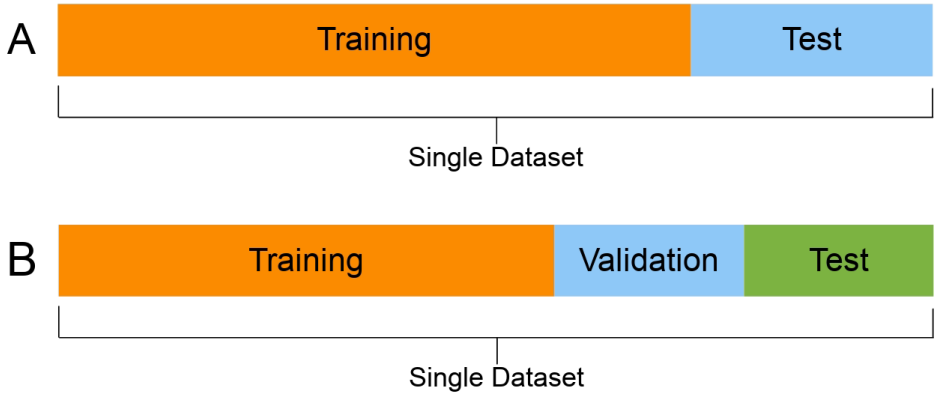
Целью работы является использование различных техник отбора признаков для обучения нейронной сети для решения задачи идентификации струй, образованных W-бозоном.

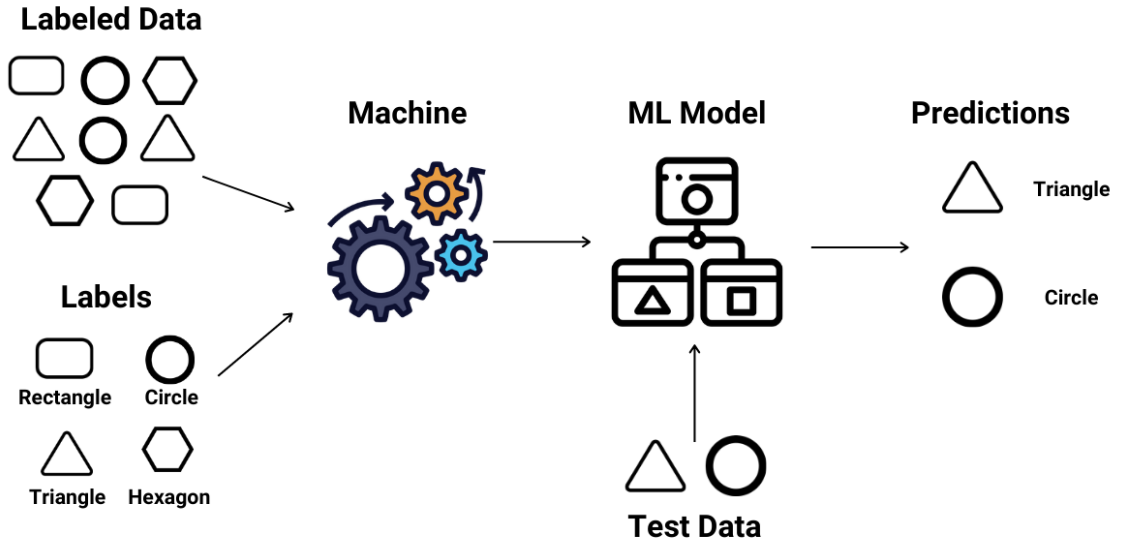
В соответствии с поставленной целью задачами данной работы были:

- Ознакомление с различными техниками отбора признаков
- Обучение и тестирование моделей
- Резюме результатов

Машинное обучение с учителем

Машинное обучение – это область прикладной математики, изучающая методы решения задач с использованием данных.





Обучающая выборка (на которой модель обучают)

Валидационная выборка (для оценки переобученности модели, для оценки ошибки прогнозирования при выборе модели, для настройки гиперпараметров и выбора лучшей модели), показывает, как может повести себя модель с новыми данными.

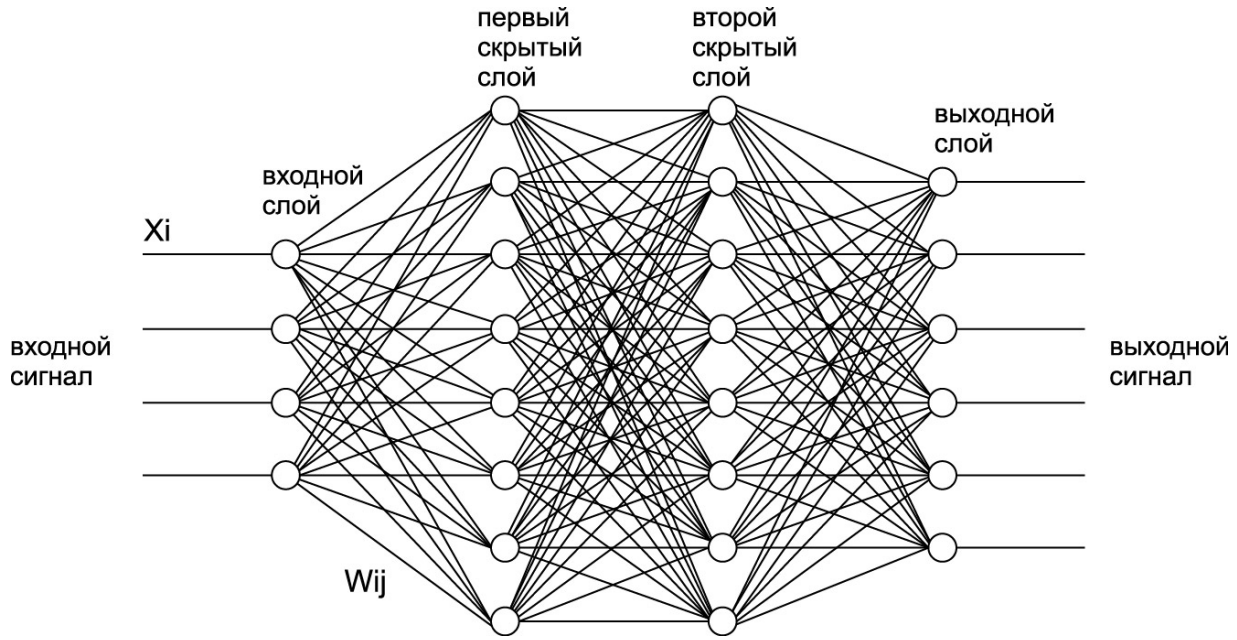
Тестовая выборка (для оценки работы готовой модели)

Используемая архитектура модели

Основные гиперпараметры MLP:

- Размер батча
- Скорость обучения
- Функция активации
- Количество нейронов/слоев
- Выбор алгоритма оптимизации

Многослойный персептрон (MLP)



Техники отбора признаков

Методы фильтрации

- Коэффициент корреляции

Метод обертки

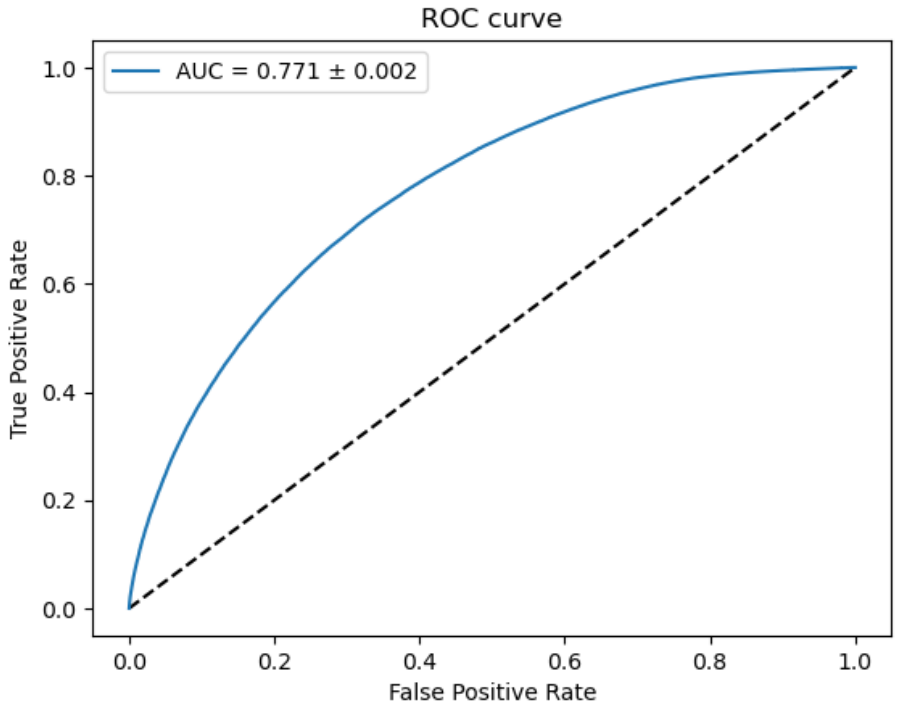
- Последовательный отбор признаков

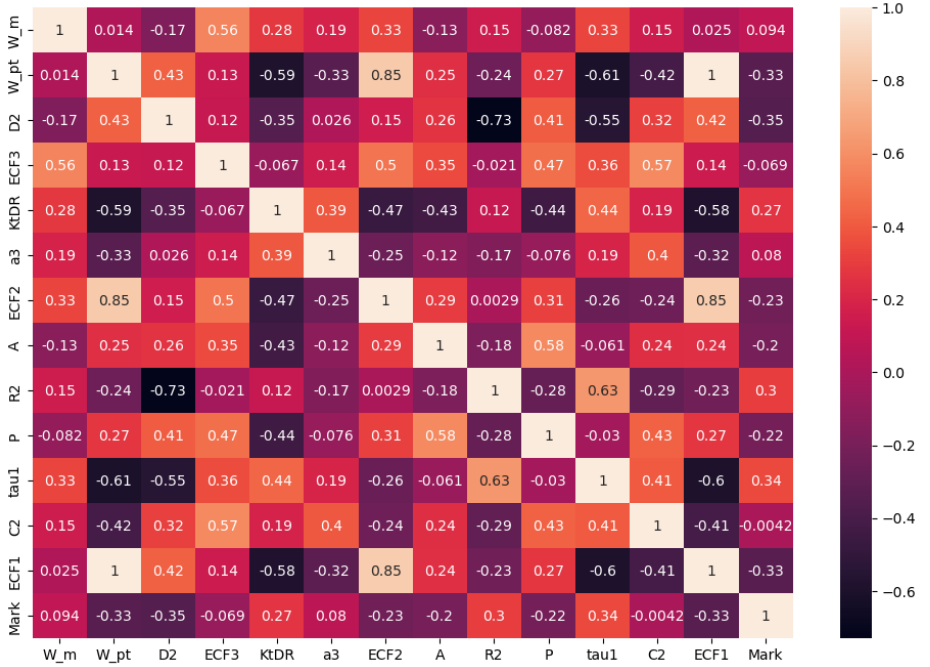
Метод с использованием деревьев решений

Метод с использованием логистической регрессии

Метод фильтрации: Корреляция Пирсона

Результаты модели:





С порогом 0.1 выявлено 9 признаков, имеющих некоторую зависимость с целевой переменной.

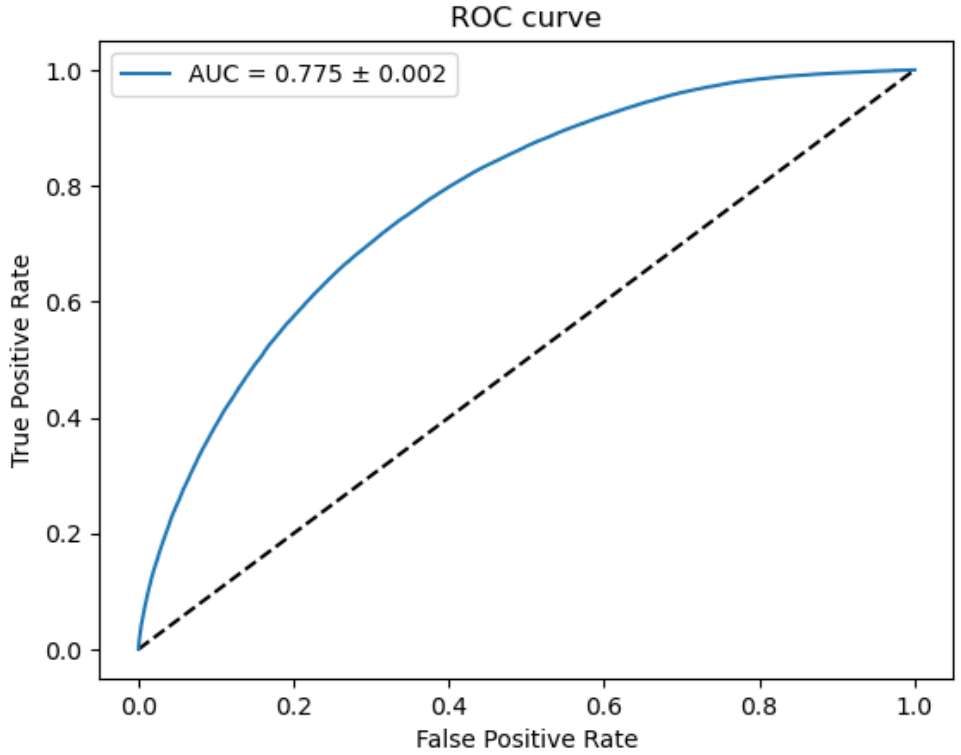
Плюсы:

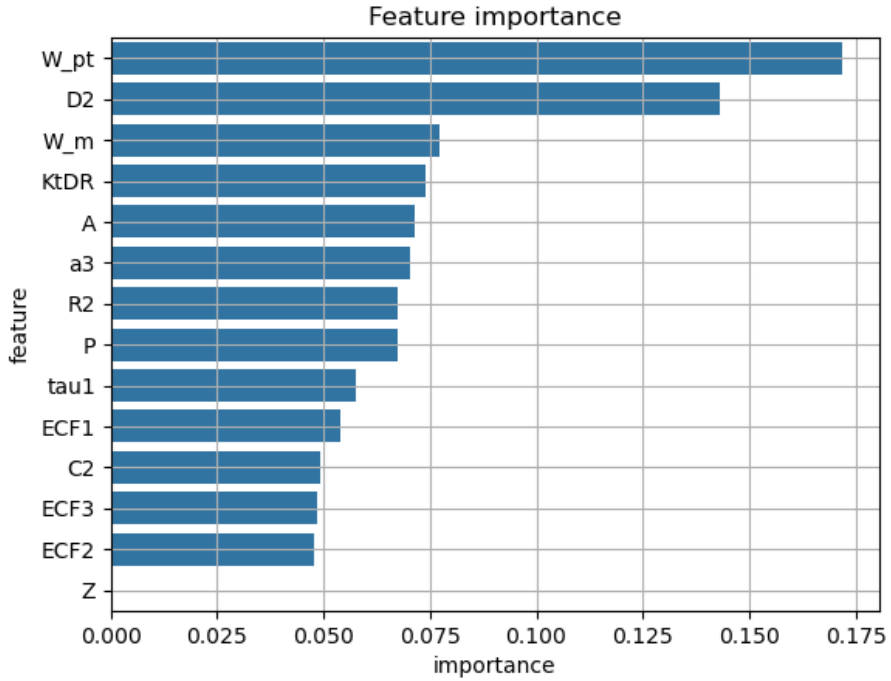
- Дает представление о том, насколько хорошо переменные связаны друг с другом.

Минусы:

- Не определяет причинно-следственную связь между любыми двумя переменными.

Метод отбора признаков с использованием дерева решений





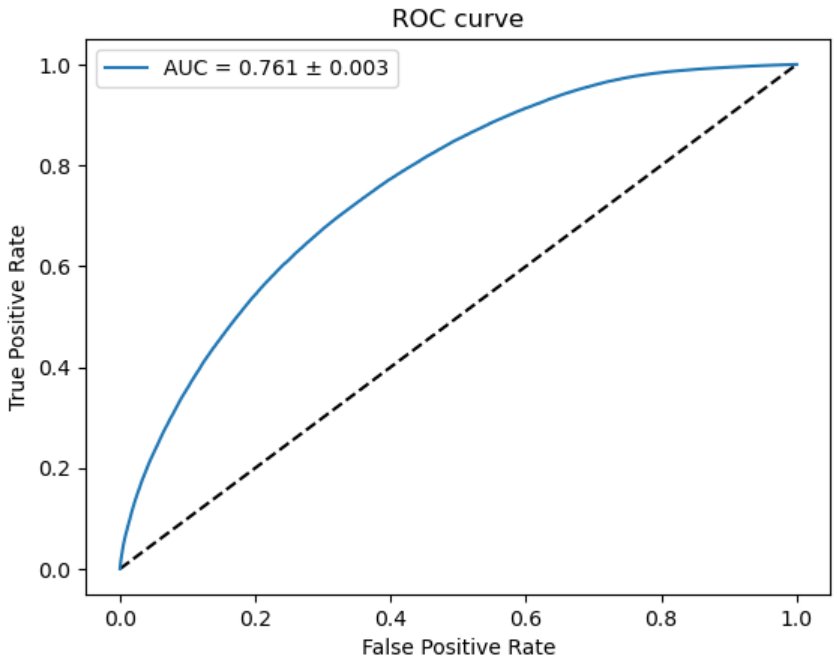
Плюсы:

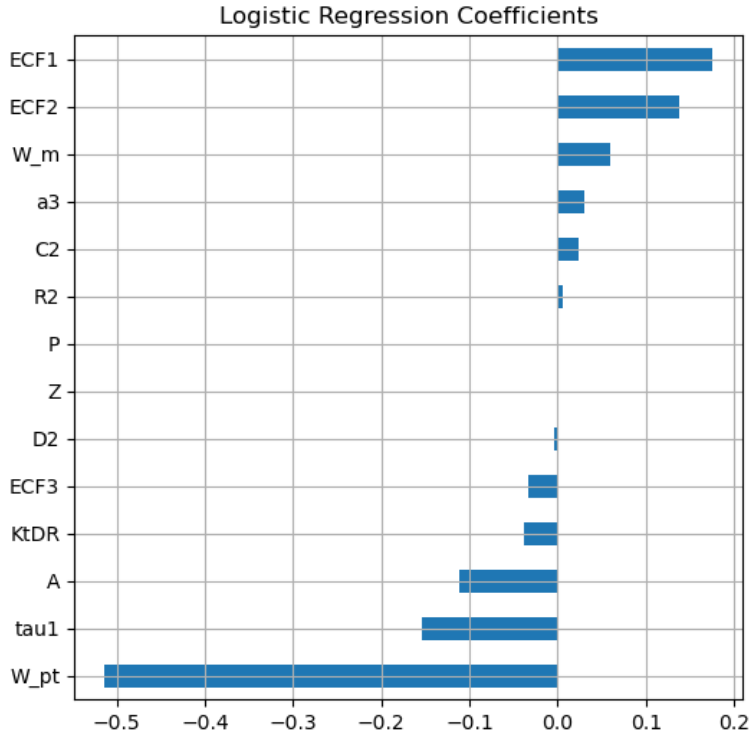
- Не требуется предобработка данных (нормализация, масштабирование, избавление от NULL)
- Простота интерпретации

Минусы:

- Расчет дерева решений требует больше ресурсов по памяти и времени
- Не устойчив к любым изменениям в данных

Метод отбора признаков с использованием логистической регрессии + LASSO





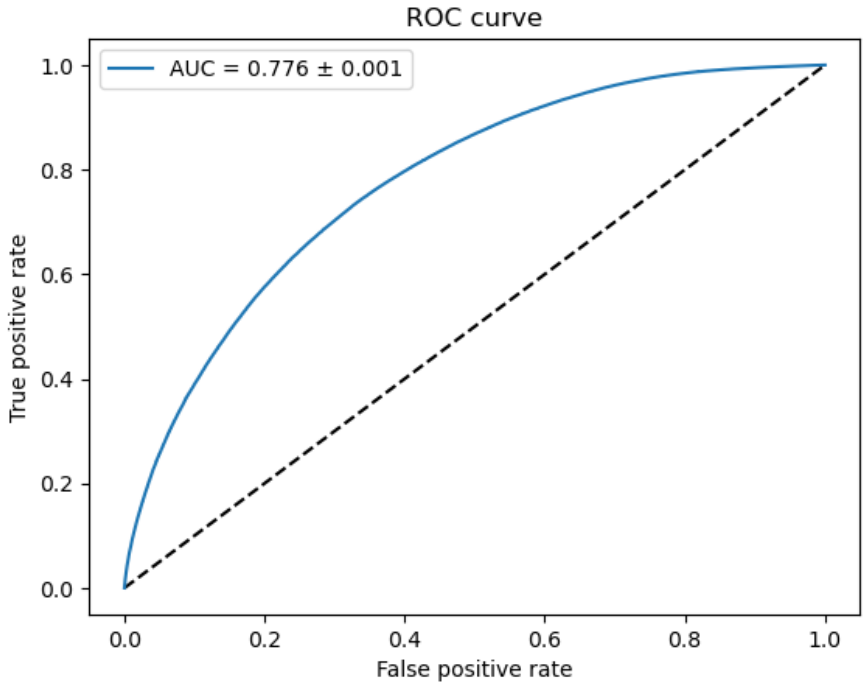
Плюсы:

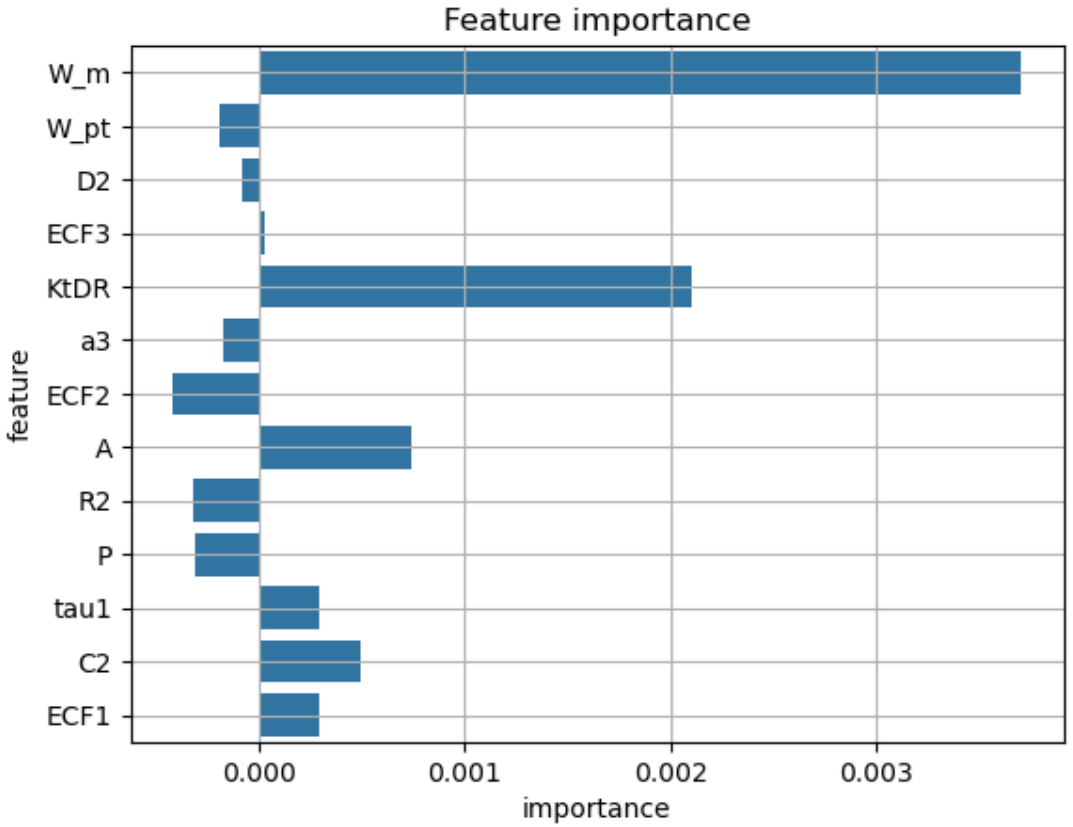
- Помимо представления меры того, насколько релевантным является размер коэффициента логистической регрессии, показывает отрицательную или положительную связь признака с целевой переменной

Минусы:

- Предполагает, что существует линейность между зависимыми и независимыми переменными, что редко бывает, поскольку данные обычно не организованы

Метод обертки: последовательный отбор признаков





Плюсы:

- Ориентирован на конкретную модель

Минусы:

- Вычислительно затратен и не оптимален

Заключение

В рамках НИР за семестр проведено ознакомление с различными методами отбора признаков для дальнейшего обучения нейронной сети.

1. Сформированы датасеты с различным набором признаков по результатам методов отбора
2. Проведено обучение и тестирование MLP для разных наборов признаков

Наилучшее значение метрики AUC показала нейронная сеть, обученная на всех сформированных признаках.

Задачи для дальнейшей работы следующие:

1. Используя наилучший набор признаков, необходимо обучить нейронную сеть на всех фоновых процессах Монте-Карло данных с учетом **веса событий**
2. Сравнить результаты работы обученной модели на **реальных** данных с работой модели на Монте-Карло симуляциях

Спасибо за внимание!