МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего

образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Отчет по заданию №4

по дисциплине

«ГЛУБИННОЕ ОБУЧЕНИЕ»

**«Задание №4 – Оценка важности признаков и визуализация в глубоком обучении»**

Выполнил

Студент гр. 932226

\_\_\_\_\_\_\_\_\_\_\_Савенкова М. М. «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2023 г

Проверил

\_\_\_\_\_\_\_\_ к.т.н. Аксёнов С.В.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2023 г.

Томск – 2023

Оглавление

[1 Индивидуальное задание 4](#_Toc154568773)

[2 Цель работы 6](#_Toc154568774)

[3 Листинг программ 7](#_Toc154568775)

[4 Анализ важности признаков 8](#_Toc154568776)

[4.1 Свёрточная модель 10](#_Toc154568777)

[4.1.1 Визуализация промежуточных состояний 11](#_Toc154568778)

[4.1.2 Визуализация фильтров сверточной нейронной сети 13](#_Toc154568779)

[4.1.3 Визуализация тепловых карт активации классов на изображениях с помощью Grad-CAM 14](#_Toc154568780)

[4.1.4 Визуализация тепловых карт активации классов на изображениях с помощью значений Шепли (image\_plot) 14](#_Toc154568781)

[4.2 Мультиклассовый классификатор 17](#_Toc154568782)

[4.2.1 Оценка важности признаков как среднее значение абсолютных значений Шепли по каждому из признаков 18](#_Toc154568783)

[4.2.2 Визуализация значений Шепли для наборов данных 19](#_Toc154568784)

[4.2.3 Зависимости между признаками 20](#_Toc154568785)

[4.3 Нейросетевой регрессор 27](#_Toc154568786)

[4.2.1 Оценка важности признаков как среднее значение абсолютных значений Шепли по каждому из признаков 27](#_Toc154568787)

[4.2.2 Визуализация значений Шепли для наборов данных 28](#_Toc154568788)

[4.2.3 Зависимости между признаками 29](#_Toc154568789)

[4.2.5 Графики частичной зависимости и индивидуального условного ожидания 33](#_Toc154568790)

[5 Выводы 36](#_Toc154568791)

# 1 Индивидуальное задание

Задачи – Получить объясняющие модели для полносвязных моделей классификации и регрессии, а также сверточной модели из задач, решенных ранее.

Для нейросетевых классификаторов и регрессора выполнить следующие задачи:

* Получить значения Шепли для примеров из наборов, используемых для построения классификатора (для каждого класса) и регрессора.
* Получить оценку важности признаков как среднее значение абсолютных значений Шепли по каждому из признаков.
* Визуализировать значения Шепли для наборов данных и полученные значения важности (summary\_plot).
* Визуализировать наиболее важные зависимости перед признаками с помощью dependence\_plot. Выбрать ряд примеров из выборок (для классификатора – по два примера для каждого класса, для регрессора – три любых примера), для которых получить объяснения с помощью decision\_plot, force\_plot, waterfall\_plot.
* Получить графики частичной зависимости и индивидуального условного ожидания в нейросетевой модели-регрессоре для признаков.

Для нейросетевой сверточной модели выполнить следующие задачи:

* Визуализировать промежуточные активации для изображений из выборки (по одному изображений из каждого класса).
* Визуализировать фильтры сверточной нейронной сети.
* Визуализировать тепловые карты активации классов на изображениях с помощью Grad-CAM и значений Шепли (image\_plot).

Сделать выводы по выполненным заданиям и привести в отчете.

# 2 Цель работы

Цель работы – получить навыки оценки важности признаков с помощью значений Шепли, визуализации активации карт признаков нейросети, признаков, извлекаемых моделью, признаков входных данных, влияющих на результат работы нейросети.

# 3 Листинг программ

Листинги программ представлены в отдельных файлах:

* CNN\_Features\_importance\_(Savenkova\_932226).ipynb
* MulClassificator\_Features\_importance\_(Savenkova\_932226).ipynb
* Regressor\_Features\_importance\_(Savenkova\_932226).ipynb

# 4 Анализ важности признаков

Значения Шепли (Shap values) – это метод для определения влияния каждого признака на предсказание модели в машинном обучении. Он основан на концепции кооперативных игр и распределения вклада (вклад каждого признака в предсказание модели). Значения Шепли позволяют определить, сколько каждый признак вносит в предсказание модели, учитывая взаимодействие с остальными признаками.

Концепция изначально была разработана для объяснения, как прибыль распределяется между игроками в кооперативных играх. В контексте машинного обучения, значения Шепли позволяют определить, в какой мере каждый признак способствует предсказанию, а также выявить взаимосвязи между признаками.

Значения Шепли можно использовать для объяснения предсказаний моделей машинного обучения и помочь исследователям и практикам понять, какие признаки наиболее влиятельны для модели, что помогает улучшить интерпретируемость модели и делать более информированные выводы.

Библиотека Shap предоставляет функционал для построения нескольких видов графиков на основании значений Шепли:

* summary\_plot — все признаки перечислены по оси y в порядке ранжирования, верхний из которых вносит наибольший вклад в прогнозы, а нижний — наименьший или нулевой вклад. Значения формы представлены по оси x. Нулевое значение не представляет никакого вклада, тогда как вклад увеличивается по мере удаления значения формы от нуля.
* dependence\_plot — графики зависимостей могут быть очень полезны при анализе важности объектов и выборе объектов. Это позволяет построить график для двух объектов, построив значения shap для одного объекта и раскрасив точки относительно другого объекта.
* decision\_plot — график принятия решения. В качестве сводного графика он дает общую картину вклада в прогнозирование. Снизу-вверх графика принятия решения значения shap кумулятивно добавляются к базовому значению модели при определении выходных значений. Можно заметить, что определенные строки, окрашенные в синий цвет, привели к окончательному значению класса 0, а остальные строки, окрашенные в красный цвет, привели к окончательному значению класса 1.
* force\_plot — графики силы подходят для построчного анализа shap-значений. График берет одну строку и показывает в порядке ранжирования, как каждый из объектов повлиял на прогноз. Чем шире блок объекта, тем больше вклад.
* waterfall\_plot — По сравнению с графиком принятия решений, waterfall\_plot представляет вклады в виде столбчатой диаграммы, что упрощает определение положительного и отрицательного влияния каждого объекта. Этот тип графика силы показывает, как значения формы каждого объекта объединяются для получения окончательного прогноза. График начинается с базового значения, а затем показывает, как каждый признак вносит вклад в общий прогноз. Эта визуализация позволяет идентифицировать признаки, которые больше всего влияют на прогнозы модели.
* partial\_dependence\_plot — показывает незначительное влияние одного или двух признаков на прогнозируемый результат модели машинного обучения.

Также для построения графиков индивидуального условного ожидания (Individual Conditional Expectation, ICE) была использована библиотека pycebox и предоставляемая ей функция ice\_plot.

4.1 Свёрточная модель

Входными данными являются изображения из набора plantdisease, расположенного по ссылке https://www.kaggle.com/datasets/emmarex/plantdisease.

Классы:

* Класс 0: Tomato\_YellowLeaf\_Curl\_virus,
* Класс 1: Tomato\_healthy.

Набор данных изображений листьев больных растений и соответствующих меток. Входные данные представлены в виде изображений, рассортированных по директориям, соответствующим классам. Каждое изображение имеет размеры 256x256 и является трёхканальным.

В качестве модели была использована модель №1 из предыдущей лабораторной работы с одним сверточным слоем и одним полносвязным.

Параметры свёрточного слоя:

* размеры входного тензора – (256,256,3);
* количество фильтров – 3;
* размер ядра каждого фильтра – (3,3);
* функция активации – relu;
* размер окна, в котором выбирается максимальное значение – (2,2).

Параметры полносвязного слоя:

* количество нейронов – 1;
* функция активации – sigmoid.

Параметры обучения:

* функция потерь – бинарная кросс-энтропия;
* оптимизатор – adam;
* метрики – BinaryAccuracy, AUC, Precision, Recall;
* взвешенные метрики – Accuracy;
* callback-функции – EarlyStopping (завершение обучения, если улучшение точности не наступит после заданного числа эпох, в нашем случае эпох 7), ModelCheckpoint (сохранение лучшей модели).

4.1.1 Визуализация промежуточных состояний

Для визуализации промежуточных состояний на вход в нейронную сеть подается изображение из тестовой выборки (рисунок 1). На выходе сверточного слоя получаем массив карт признаков, которые можно визуализировать.

|  |  |
| --- | --- |
| C:\Users\User\Downloads\0172fce9-ee7c-4bb8-9a46-7fc0050438b5___UF.GRC_YLCV_Lab 03161.JPG | C:\Users\User\Downloads\02b4afdf-e1de-4c0e-a38d-3f19afeb9ea9___RS_HL 0493.JPG |
| Рисунок 1 — Изображения класса 0 и класса 1 для примера | |

На рисунках 2 и 3 представлены промежуточные состояния для изображений разных классов.

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\1а.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\1б.png |
| Рисунок 2 — Промежуточные состояния класса 0 |

На рисунке для класса 0, то есть для томатов, зараженных вирусом, видно, что на втором изображении выделился имеющий более светлую окраску край листа, а на третьем – более неровная текстура.

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\2а.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\2б.png |
| Рисунок 3 — Промежуточные состояния класса 0 |

4.1.2 Визуализация фильтров сверточной нейронной сети

Визуализация фильтров сверточной нейронной сети позволяет понять, какие признаки или шаблоны изображений фильтры извлекают в процессе обучения.

На рисунке 4 проиллюстрированы веса фильтров, которые показывают форму и структуру фильтров. Яркость пикселей соответствует весу.

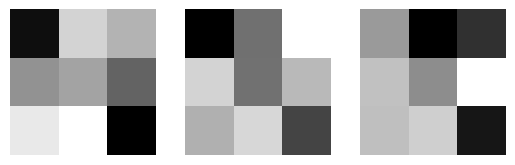


Рисунок 4 — Фильтры

4.1.3 Визуализация тепловых карт активации классов на изображениях с помощью Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) – это техника визуализации в машинном обучении, которая позволяет определить, какие части изображения были наиболее значимыми для принятия решения моделью глубокого обучения, что помогает интерпретировать ее работу. Она основана на анализе градиентов и активаций внутри нейросети.

На рисунке 5 представлены тепловые карты для изображений двух классов.

|  |  |
| --- | --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\4.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\5.png |
| Рисунок 5 — Тепловые карты для класса 0 и класса 1 | |

4.1.4 Визуализация тепловых карт активации классов на изображениях с помощью значений Шепли (image\_plot)

Визуализация тепловых карт активации классов на изображениях с помощью значений Шепли может помочь проинтерпретировать, какие области изображения больше всего влияют на принятие решения классификатором.

Значения Шепли используются для определения важности каждого пикселя или области изображения для принятия решения классификатором. Чем выше значение Шепли, тем важнее пиксель или область для классификации изображения.

На рисунках 6 и 7 представлены тепловые карты для выбранных изображений.

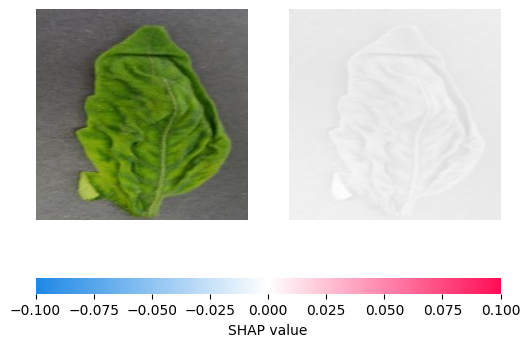


Рисунок 6 — Тепловая карта для класса 0

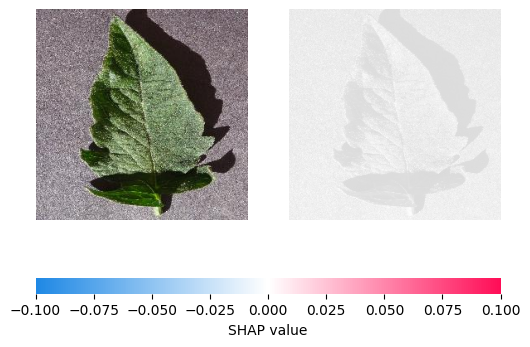


Рисунок 7 — Тепловая карта для класса 1

4.2 Мультиклассовый классификатор

Набор данных содержит 2126 записей признаков, извлеченных из результатов кардиотокограмм, которые затем были классифицированы тремя экспертами-акушерами на 3 класса:

* Normal (Нормальные)
* Suspect (Подозрительные)
* Pathological (Патологические)

Набор данных содержит следующие признаки:

* baseline\_value ‒ исходная частота сердечных сокращений плода;
* accelerations ‒ количество ускорений в секунду;
* fetal\_movement ‒ количество движений плода в секунду;
* uterine\_contractions ‒ количество сокращений матки в секунду;
* light\_decelerations ‒ количество LDS в секунду;
* severe\_deceleration ‒ количество Sds в секунду;
* prolongued\_deceleration ‒ количество PDs в секунду;
* abnormal\_short\_term\_variatability ‒ процент времени с ненормально коротким периодом изменчивости;
* mean\_value\_of\_short\_term\_variability ‒ среднее значение краткосрочной изменчивости;
* percentage\_of\_time\_with\_abnormal\_long\_term\_variability ‒ процент времени с аномальной долгосрочной изменчивостью;
* mean\_value\_of\_long\_term\_variability ‒ среднее значение длительной вариабельности;
* histogram\_width ‒ ширина гистограммы с использованием всех значений из записи;
* histogram\_min ‒ минимальное значение гистограммы;
* histogram\_max ‒ максимальное значение гистограммы;
* histogram\_number\_of\_peaks ‒ количество пиков в гистограмме;
* histogram\_number\_of\_zeroes ‒ количество нулей в гистограмме;
* histogram\_mode ‒ режим гистограммы;
* histogram\_mean ‒ среднее значение гистограммы;
* histogram\_median ‒ медиана гистограммы;
* histogram\_variance ‒ дисперсия гистограммы;
* histogram\_tendency ‒ тенденция гистограммы;
* fetal\_health ‒ здоровье плода.

4.2.1 Оценка важности признаков как среднее значение абсолютных значений Шепли по каждому из признаков

Средние значения абсолютных значений Шепли по каждому признаку показывают важность каждого признака в объяснении прогнозных моделей. Чем выше среднее значение абсолютного значения Шепли для конкретного признака, тем больше вклад этого признака в прогноз.

На рисунке 8 приведен результат вычисления средних значений абсолютных значений Шепли по каждому классу.

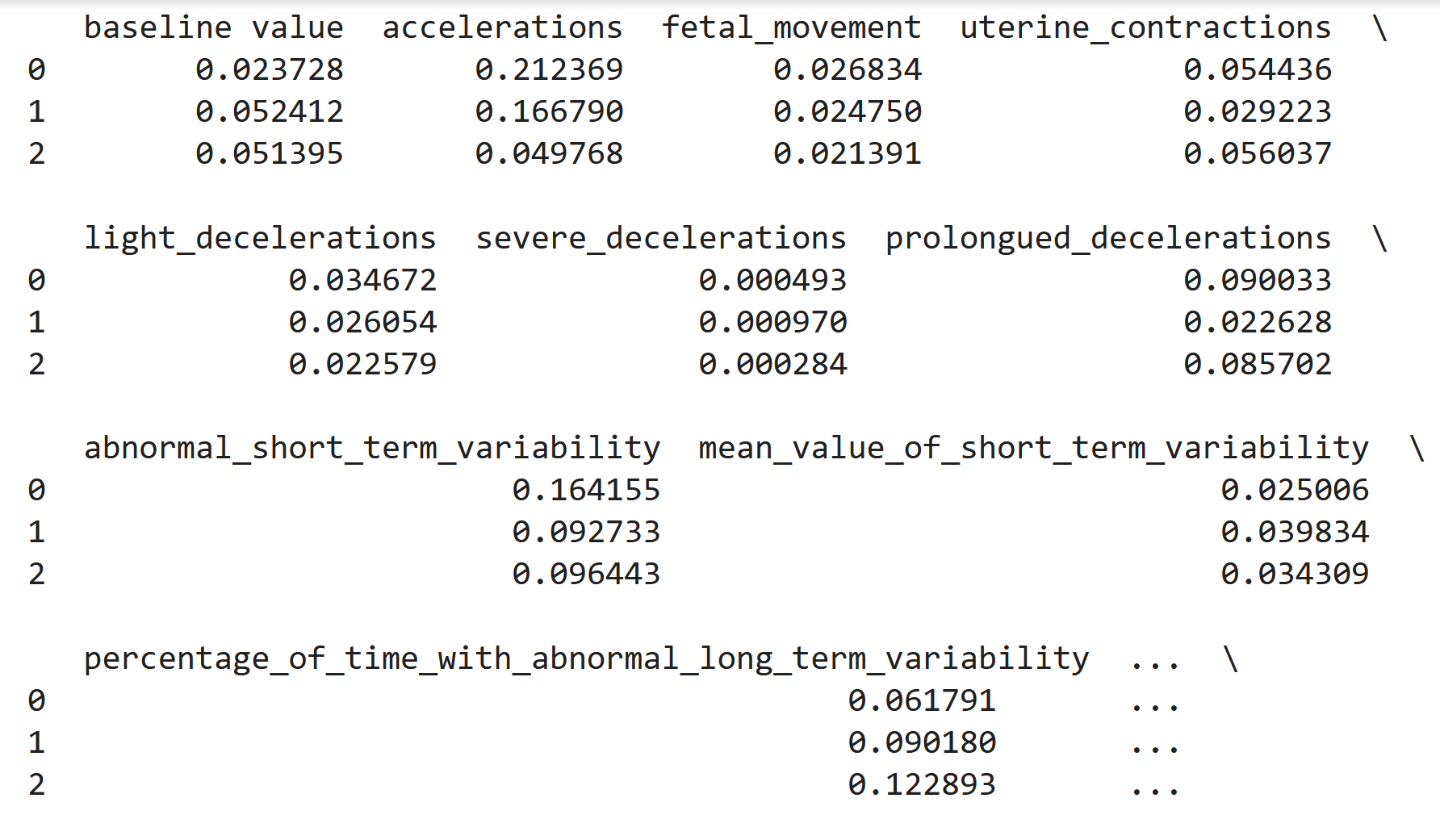


Рисунок 8 — Результат вычислений абсолютных средних значений по каждому классу

4.2.2 Визуализация значений Шепли для наборов данных

На рисунке 9 представлен summary\_plot. Как видим, наиболее важным признаком является accelerations, а наименьший вклад внес признак histogram\_min.

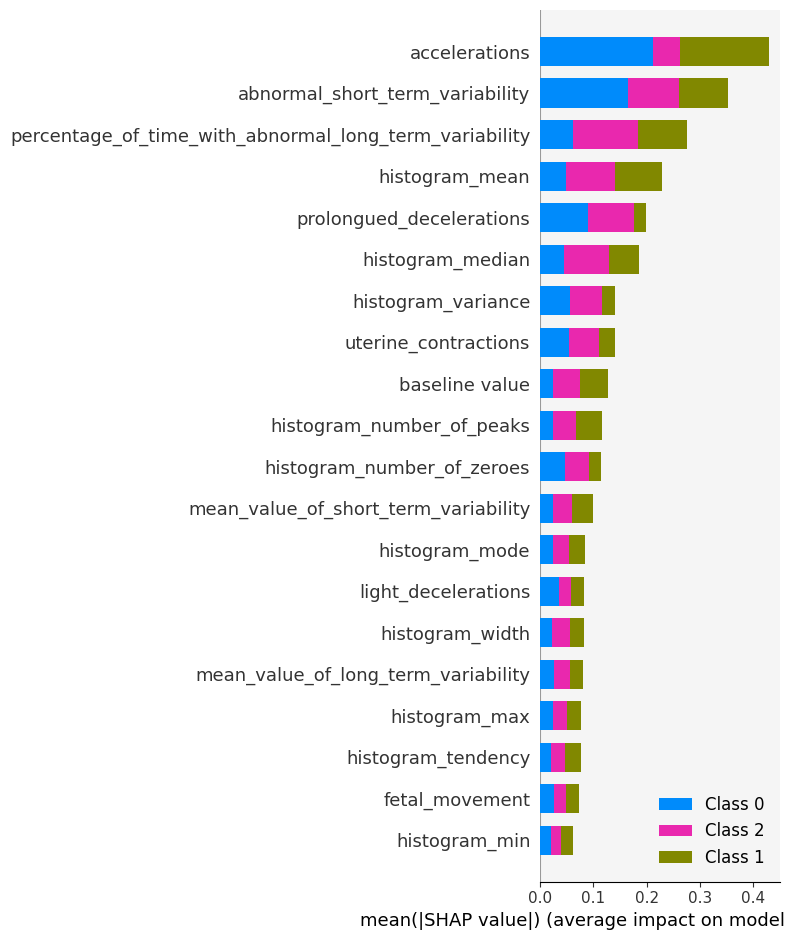


Рисунок 9 — Summary plot

4.2.3 Зависимости между признаками

На рисунках 10 и 11 приведены графики dependence\_plot наиболее важных признаков для каждого класса. Синий цвет соответствует классу 0, красный – классу 1. Фиолетовые точки имеют наименьшую важность.

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\111.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\222.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\333.png |
| Рисунок 10 – Графики зависимости для accelerations |

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\bv_0.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\bv_1.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\bv_2.png |
| Рисунок 11 – Графики зависимости для baseline value |

Для каждого класса было выбрано по два примера и получены объяснения с помощью decision\_plot, force\_plot, waterfall\_plot. На рисунках 12, 13, 14 приведены графики для первого примера, принадлежащему классу 0 (нормальные наблюдения), который ошибочно был классифицирован как класс 1 (подозрительные).

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\1.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\2.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\3.png |
| Рисунок 12 – Графики для класса 0 (первый пример) |

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\4.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\5.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\6.png |
| Рисунок 13 – Графики для класса 1 (первый пример) |

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\7.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\8.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\9.png |
| Рисунок 14 – Графики для класса 2 (первый пример) |

4.3 Нейросетевой регрессор

В данной работе был использован набор данных DS\_2019\_public.csv, содержащий информацию о потреблении ресурсов домохозяйством. Описание атрибутов содержится в файле recs2009\_public\_codebook.xlsx. Всего датасет содержит 121 признак.

В качестве целевой переменной была выбрана TOTALBTU - общее потребление (в тысячах BTU).

4.2.1 Оценка важности признаков как среднее значение абсолютных значений Шепли по каждому из признаков

Значения Шепли в задаче регрессии можно интерпретировать как вклад каждого признака в прогноз модели. Они показывают, насколько в среднем меняется прогноз модели, когда к нему добавляется конкретный признак, учитывая все возможные комбинации остальных признаков.

Значение Шепли равное нулю означает, что признак не вносит вклад в прогноз модели, значение Шепли больше нуля указывает на положительный вклад признака, а значение Шепли меньше нуля — на отрицательный вклад признака.

На рисунке 15 приведен результат вычисления средних значений абсолютных значений Шепли по каждому классу.

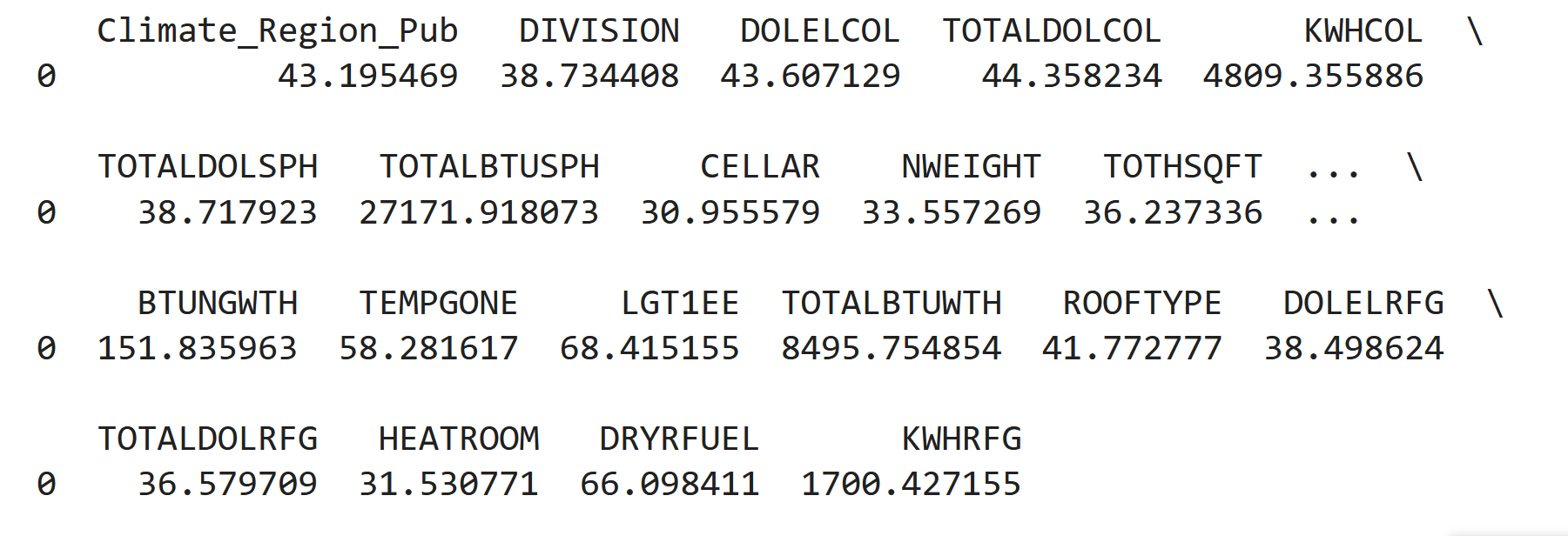


Рисунок 15 — Результат вычисления средних значений абсолютных значений Шепли по каждому классу

4.2.2 Визуализация значений Шепли для наборов данных

По графику summary\_plot (рисунок 16) можно судить о том, что наибольший вклад внесли признаки KWH и COOLTYPE.

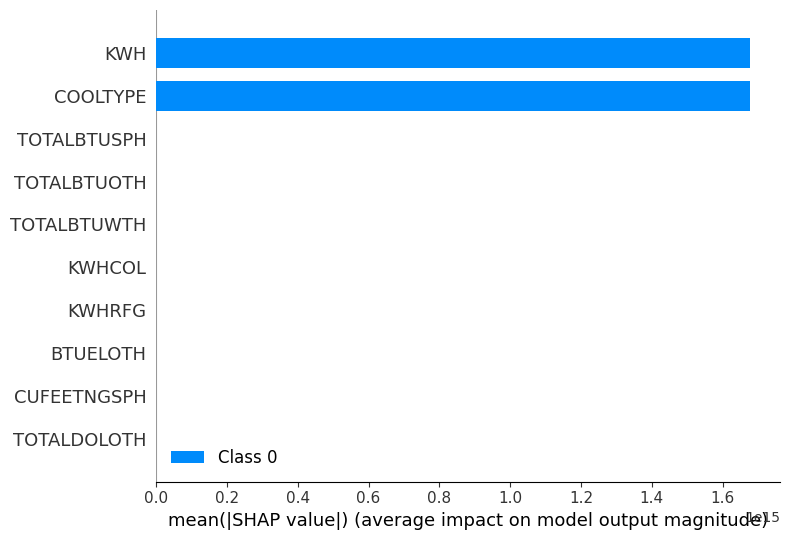


Рисунок 16 — Summary plot

4.2.3 Зависимости между признаками

На рисунке 17 приведен dependence\_plot для восьми наиболее важных признаков.

|  |  |
| --- | --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr1.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr2.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr3.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr4.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr5.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr6.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr7.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\rr8.png |
| Рисунок 17 — Dependence plot | |

В соответствии с заданием для трех примеров были получены объяснения с помощью decision\_plot, force\_plot, waterfall\_plot (рисунок 18а, 18б, 18в).

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\1.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\2.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\3.png |
| Рисунок 18а — Первый пример |

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\4.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\5.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\6.png |
| Рисунок 18б — Второй пример |

|  |
| --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\7.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\8.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\9.png |
| Рисунок 18в — Второй пример |

4.2.5 Графики частичной зависимости и индивидуального условного ожидания

График частичной зависимости (PDP plot) показывает незначительный эффект, который оказывают одна или две функции на прогнозируемый результат модели машинного обучения. График частичной зависимости может показать, является ли взаимосвязь между целью и объектом линейной, монотонной или более сложной. Например, при применении к модели линейной регрессии графики частичной зависимости всегда показывают линейную взаимосвязь.

На рисунке 19 приведены графики частичной зависимости для 10 наиболее важных признаков. На последних четырех графиках зависимость не линейная.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d1.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d2.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d3.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d4.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d5.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d6.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d7.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d8.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d9.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\d10.png |
| Рисунок 19 — Графики частичной зависимости | | | | |

На рисунке 20 приведен график частичной зависимости для двух признаков – KWH и COOLTYPE. Более яркие цвета на графике обозначают более высокую важность признака.

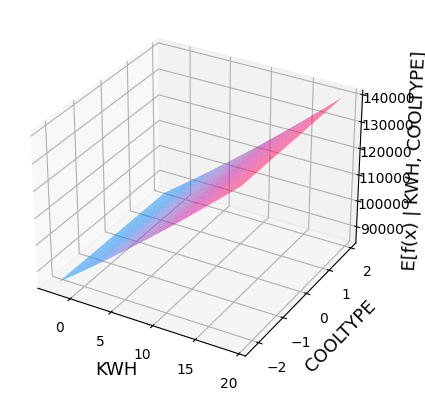


Рисунок 20 — График частичной зависимости признаков KWH и COOLTYPE

Кривая ICE — это графическое представление взаимосвязи между отдельным признаком и прогнозами модели для конкретной точки данных. На графике ICE несколько кривых ICE отображаются вместе, причем каждая кривая представляет отдельную точку данных. График позволяет визуализировать, как изменяются прогнозы модели по мере изменения значения признака для каждой отдельной точки данных, показывая влияние признака на прогнозы модели.

На рисунке 21 представлены графики ICE для нескольких наиболее важных признаков.

|  |  |
| --- | --- |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\i1.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\i2.png |
| C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\i3.png | C:\MARINA\STUDY\Магистратура\семестр 3\глубинное обучение\задание 4\i4.png |
| Рисунок 21 — ICE для наиболее важных признаков | |

# 5 Выводы

В данной лабораторной работе были получены объясняющие модели для полносвязных моделей классификации и регрессии, а также сверточной модели из задач, решенных ранее. Использованные в работе методы позволяют определить важность признаков и получить промежуточные результаты работы нейронной сети, что можно использовать для проверки решений, принимаемых моделью, чтобы определить, какие факторы приводят к тому или иному прогнозу.