Оглавление

[ВВЕДЕНИЕ 2](#_Toc47102654)

[Используемые термины и понятия 3](#_Toc47102655)

[Обзор данных экспериментов 3](#_Toc47102656)

[Сшивка 1 – Первый тест на температуру 3](#_Toc47102657)

[Сшивка 2 – Второй тест на температуру 4](#_Toc47102658)

[Первый тест с газом 5](#_Toc47102659)

[Второй тест с газом 6](#_Toc47102660)

[Саморазогрев без газов – первый тест 7](#_Toc47102661)

[Саморазогрев без газов – второй тест 8](#_Toc47102662)

[Тест на открытом воздухе всю ночь 9](#_Toc47102663)

[Анализ и фильтрация данных 10](#_Toc47102664)

[Сшивки 1 и 2 – Температурные тесты 10](#_Toc47102665)

[Тесты на газах 1 и 2 12](#_Toc47102666)

[Тесты на саморазогрев 1 и 2 17](#_Toc47102667)

[Тест на открытом воздухе всю ночь 18](#_Toc47102668)

[Разделение данных на обучающую и тестовую выборки 19](#_Toc47102669)

[Построение и испытание нейронных сетей 19](#_Toc47102670)

[Итоговая проверка моделей 23](#_Toc47102671)

[Заключение по результатам 24](#_Toc47102672)

[Список литературы 24](#_Toc47102673)

# ВВЕДЕНИЕ

Целью данной работы является анализ данных, полученных с LED датчика

MDS-4-2 в ходе ряда экспериментов, и построение моделей корректировки показаний датчика по содержанию CH4 в смеси газов.

Для достижения лучшей точности показаний датчика были поставлены следующие задачи:

- изучить результаты экспериментов и выделить информативные данные;

- изучить структуру данных, проанализировать выявленные закономерности, отфильтровать шумовые данные, подготовить тренировочную и тестовую выборки;

- оценить эффективность простых сглаживающих методов;

- оценить эффективность линейных методов корректировки;

- построить нейронную сеть и оценить её эффективность;

- построить итоговую корректирующую модель.

# Используемые термины и понятия

Um – амплитуда импульсов напряжения на выходе предусилителя измерительного фотодиода; измерительный сигнал.

Ur – амплитуда импульсов напряжения на выходе предусилителя эталонного фотодиода; опорный сигнал.

Ud – амплитуда импульсов напряжения светодиода, измеренная при рабочем токе; точно коррелирует с температурой.

С – значение доли метана в смеси газов в ppm.

С\_target – целевое значение доли метана в смеси газов в ppm.

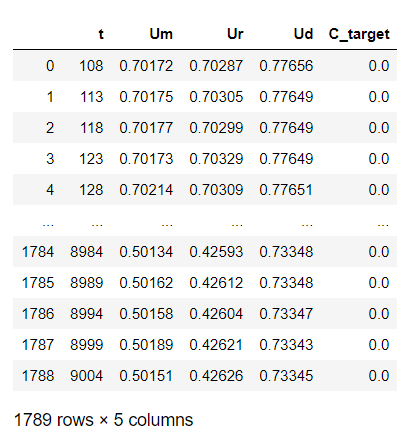
# Обзор данных экспериментов

## Сшивка 1 – Первый тест на температуру

Описание: прогон сенсора в термокамере под азотом

Ход эксперимента:

* включение сенсора, подача азота, прогрев, постановка на запись
* охлаждение термокамеры до +10 град, выдержка 30 мин
* с 3119 сек включение режима +40 град с шагом около 5 град, чтобы замедлить скорость нагрева
* с 7194 сек полное выключение термокамеры, сенсор остывает на записи под азотом



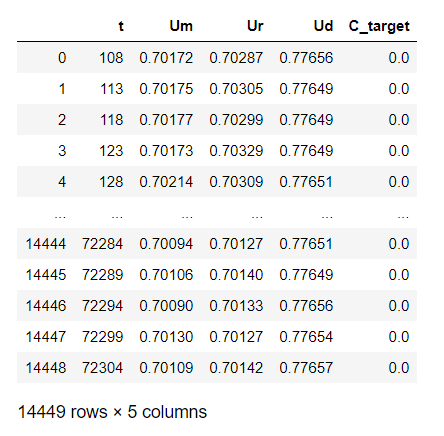
Сшивка 1 (stitch 1 view)

## Сшивка 2 – Второй тест на температуру

Описание: прогон сенсора в термокамере под азотом, остывание записывалось всю ночь

Ход эксперимента:

* включение сенсора, подача азота, прогрев, постановка на запись
* охлаждение термокамеры до +10 град, выдержка 30 мин
* с 3119 сек включение режима +40 град с шагом около 5 град, чтобы замедлить скорость нагрева
* с 7194 сек полное выключение термокамеры, сенсор остывает на записи под азотом



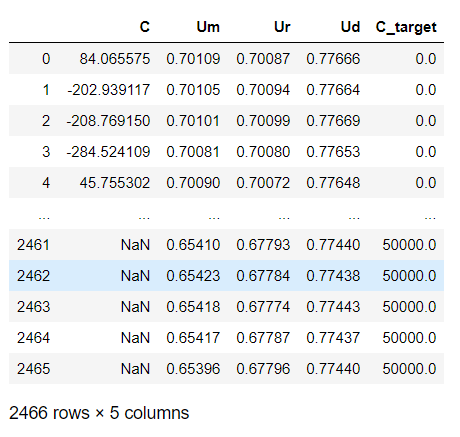
Сшивка 2 (stitch 2 view)

## Первый тест с газом

Описание: последовательная продувка газами по 25 минут

Ход эксперимента:

* азот с 1592 сек
* 0,05% с 3042 сек
* 0,20% с 4502 сек
* 0,50% с 6032 сек
* 1% с 7522 сек
* 2% с 8962 сек
* 3% с 10582 сек
* 5% с 12087 сек



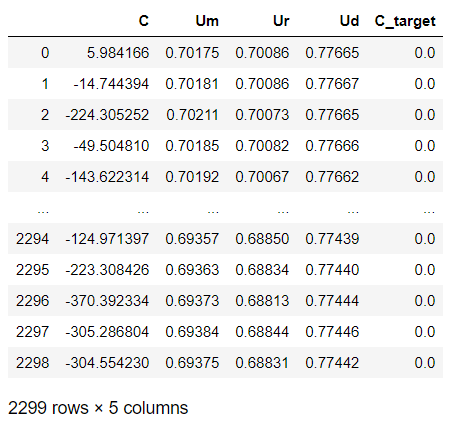
Тест с газом 1 (gases 1 view)

## Второй тест с газом

Описание: последовательная продувка газами

Ход эксперимента:

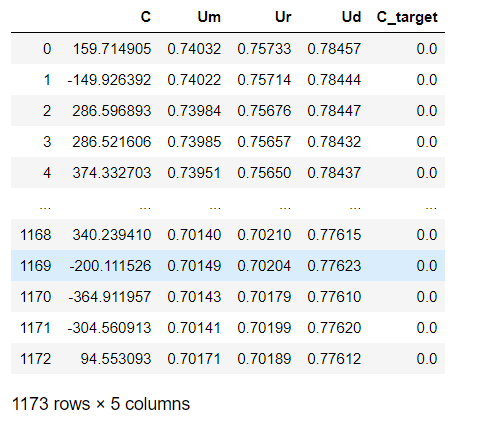
* азот с 70 сек (13:46-14:50)
* 0,05% с 3730 сек (14:50-15:10)
* 0,20% с 5100 сек (15:12-15:30)
* 2% с 6300сек (15:32-15:52)
* 3% с 7475сек (15:52-16:12)
* 5% с 8695сек (16:12-16:32)
* азот с 9950сек (16:33-16:53)



Тест с газом 2 (gases 2 view)

## Саморазогрев без газов – первый тест

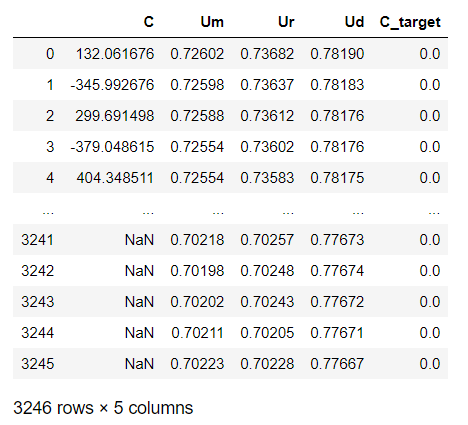
Описание: сенсор испытываплся на саморазогрев без газов, самопроизвольно выключился спустя час



Саморазогрев 1 (temperature 1 view)

## Саморазогрев без газов – второй тест

Описание: сенсор 36 испытывался на саморазогрев и зависимость от внешней температуры без газов

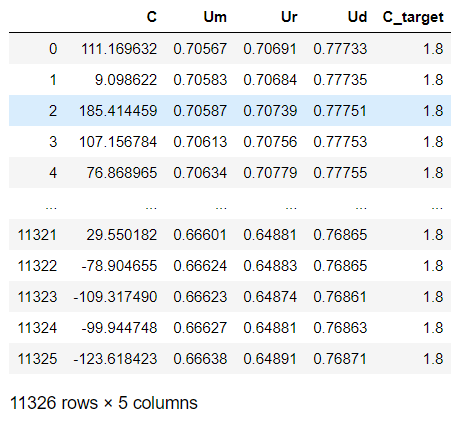


Саморазогрев 2 (temperature 2 view)

## 

## Тест на открытом воздухе всю ночь

Описание: сенсор 36 простоял в открытом виде (не в коробке) на воздухе всю ночь

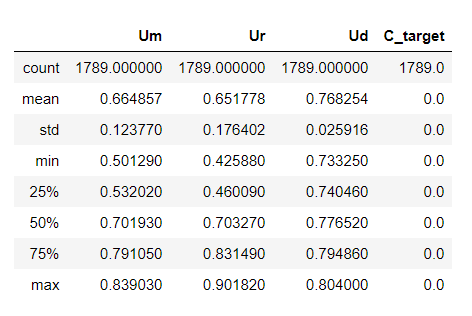


Всю ночь (all night view)

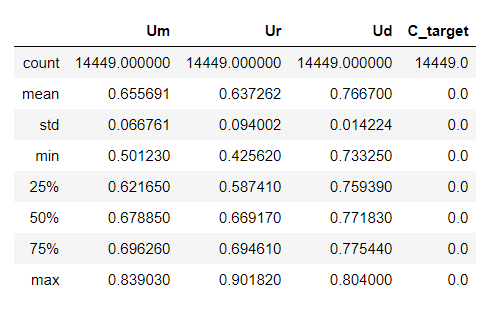
# Анализ и фильтрация данных

## Сшивки 1 и 2 – Температурные тесты

Рассмотрим стандартные выборочные характеристики выборок:

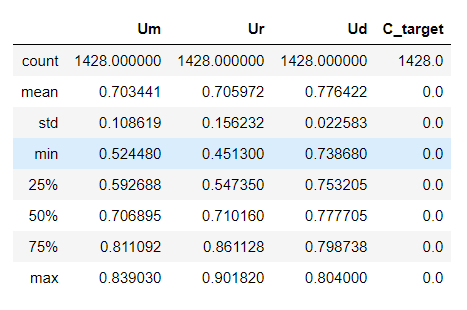


Характеристики Сшивки 1 (Stitch 1 describe)

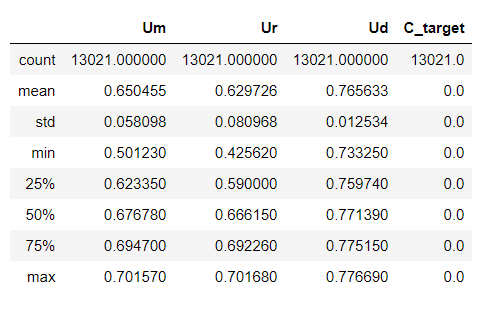


Характеристики Сшивки 2 (Stitch 2 describe)

Заметим, что стандартное отклонение на Сшивке 2 практически в два раза ниже чем на Сшивке 1, значения интерквартилей также значительно ближе к медиане. По описанию экспериментов мы видим, что в начальной части они проводились одинаково, но во втором эксперименте после 7200 секунды начался процесс остывания. Очевидно, что он проходил до какого-то стационарного значения температуры, после чего датчик мог находиться в этом состоянии большую часть времени, что отражается на средних характеристиках.



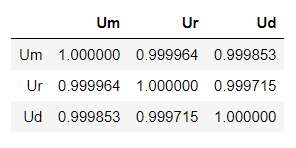
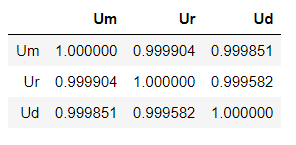
Характеристики Сшивки 2 до остывания (Stitch 2 time < 7200)



Характеристики Сшивки 2 в процессе остывания (Stitch 2 time > 7200)

Как видим из данных, главной причиной малой дисперсии во втором эксперименте, как и ожидалось является остывание. Эти данные для наших целей представляются бесполезными, так как это огромное количество объектов выборки (относительно информативных) с крайне низким разнообразием температурных значений ( что является приоритетом, так как именно температура сильно влияет на ошибку сенсора) и нулевым таргетом, что очевидно может сильно помешать, если мы хотим обучаться на этих данных. Поэтому мы можем отбросить большую часть объектов из остывания, и не потерять ценной информации.

Теперь рассмотрим значения корреляции признаков:



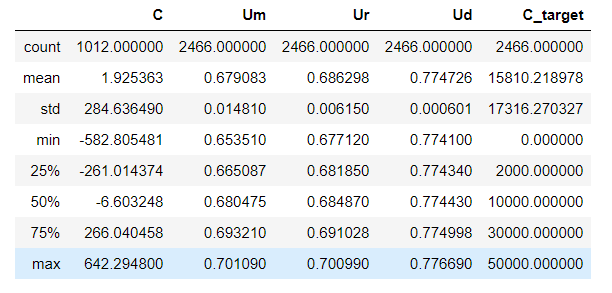
Корреляция признаков Сшивки 1 и 2 (Stitch 1-2 correlation)

Видим, что признаки имеют очень сильно выраженную линейную коррелированность друг с другом, из чего следует, что линейная модель теоретически может хорошо выразить влияние температуры на определяющие признаки, и соответственно нивелировать его.

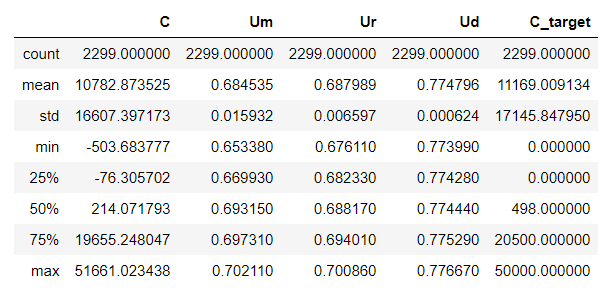
У данных выборок есть два существенных недостатка: первый – испытания проводились только на чистом азоте, соответственно наши модели просто не будут знать поведение признаков при различных значениях содержания метана на экстремальных температурных значениях, также нулевые значения таргетов на этих объектах способствует подстраиванию моделей под выборку при её высокой доле в обучающей выборке ( вместо выявления каких-то зависимостей, выборка будет поощрять зануление выходного результата); второй – на выборках отсутствуют показания датчика по значению С, что ограничивает нас в использовании данных на тестовых испытаниях ( мы можем лишь посчитать ошибку предсказаний наших моделей относительно таргета, но мы не можем посчитать эффективность корректирования показаний датчика).

## Тесты на газах 1 и 2

Рассмотрим стандартные выборочные характеристики выборок:



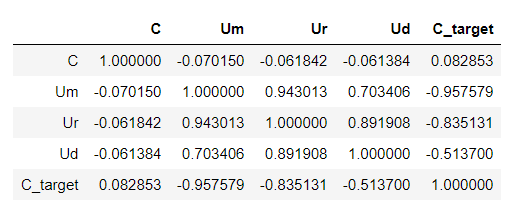
Характеристики теста газов 1 (gases 1 describe)



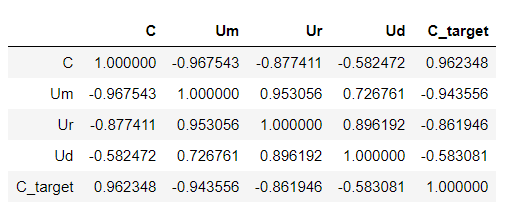
Характеристики теста газов 2 (gases 2 describe)

Можем сразу заметить, что выборки сильно разняться относительно характеристик по показаниям значения С, при относительной схожести характеристик по другим параметрам, при чем значения С в первом тесте с газами очевидно неверны (максимальное значение 642 ppm при таргете в 5000), также стоит обратить внимание на большее количество объектов с нулевым таргетом во втором тесте (нижний квартиль таргета в выборке равен нулю).

Рассмотрим значения корреляции признаков:



Корреляция признаков теста газов 1 (gases 1 correlation)



Корреляция признаков теста газов 2 (gases 2 correlation)

Видим, что при достаточно высокой коррелированности основных параметров со значением таргета, какая-либо корреляция с параметром С в первом испытании отсутствует (в отличие от второго испытания), что подтверждает наше предположение, что значения С в первом тесте являются шумом, что делает эту выборку непригодной для сравнения качества корректировки.

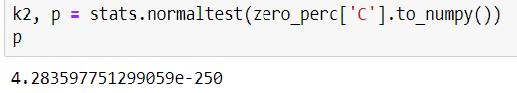
Так как процесс подачи газа в экспериментах непрерывен, то на объектах, находящихся на переходах между газовыми смесями, будет происходить лаг, вызванный диффузией газов, между реальным значением процента метана и указанным таргетом, до момента пока смесь устоится. Подобные объекты в нашей выборке будут являться выбросами, и чтобы они не мешали на обучении их необходимо отфильтровать.

Попробуем отфильтровать данные по признаку Um, так как он второй после С по уровню корреляции с таргетом (на первом тесте значение С мы использовать не можем).

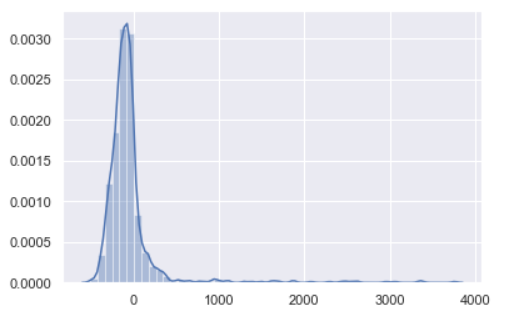
Выбросы будем классифицировать по стандартному правилу - значения, лежащие за краями статистически значимой выборки (Разность первого квартиля и полутора межквартильных расстояний; сумма третьего квартиля и полутора межквартильных расстояний).

После фильтрации по Um количество объектов уменьшилось на **8%** в первой выборке и на **7%** во второй, при исходных размерах выборки **2466** и **2299** соответственно.

Оценим качество фильтрации, рассмотрев значения показателя С на отфильтрованной выборке второго теста. При обзоре данных можно заметить, что значение С на нулевом таргете значительно отклоняется от нуля как в положительную, так и в отрицательную сторону. Известно, что любой измерительный прибор обладает некоторой естественной ошибкой в силу технических барьеров. В таком случае можно предположить, что данная ошибка будет несмещённой нормально распределённой. Проверим эту гипотезу на данных с нулевым таргетом.



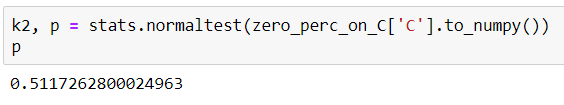
p-value практически нулевое, это означает, что данные точно не принадлежат нормальному распределению. Давайте посмотрим на их график.



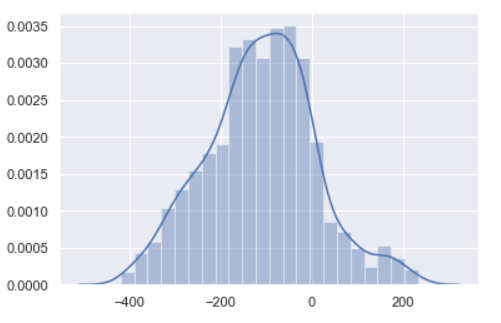
Распределение значений С на нулевом таргете

Очевидно, что подобный результат мы получили из-за недостаточной жесткости фильтрации. Подобное объяснимо малым абсолютным значением дисперсии параметра Um. Проверим эту теорию отфильтровав данные второго теста по значениям С.

После фильтрации по С количество объектов уменьшилось на **11,5%** .

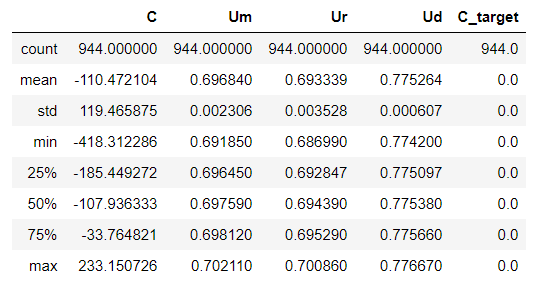


После фильтрации по С мы получили уровень p-value, достаточно высокий чтобы принять теорию о нормальности ошибки.



Распределение значений С на нулевом таргете после фильтрации по С

Рассмотрим выборочные характеристики распределения:



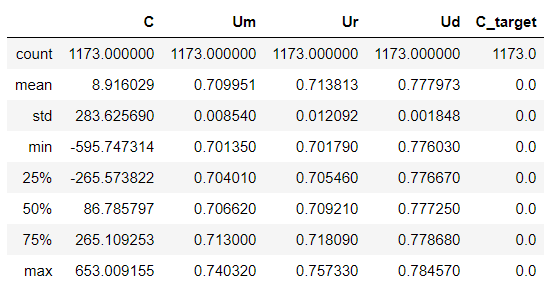
Характеристики ошибки при нулевом таргете (zeros describe)

По графику и из таблицы видно, что ошибка датчика является смещённой, что естественно негативно сказывается на точности выдаваемого датчиком результата.

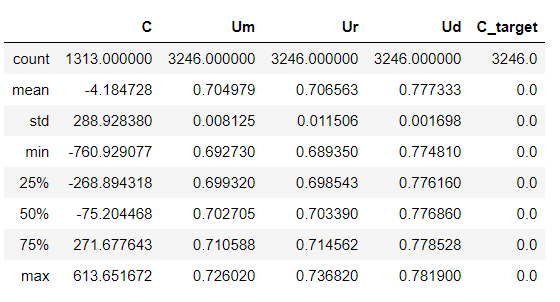
В целом, мы значительно улучшили качество наших данных, отфильтровав их по С, к несчастью данный подход не применим к данным из первой выборки.

## Тесты на саморазогрев 1 и 2

Рассмотрим стандартные выборочные характеристики выборок:

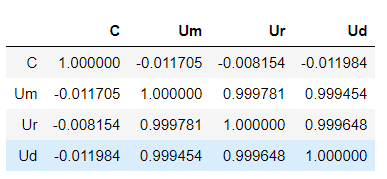
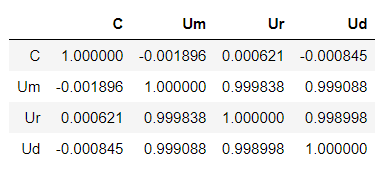


Саморазогрев 1 (temperature 1 describe )



Саморазогрев 2 (temperature 2 describe )

Рассмотрим значения корреляции признаков:

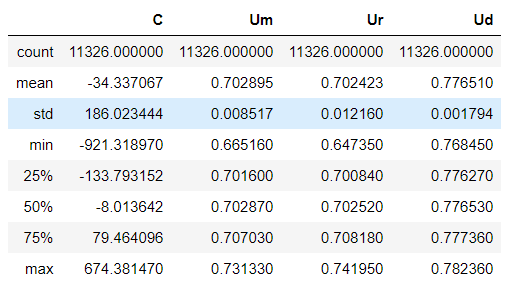


Корреляция признаков Саморазогрев 1 и 2 (temperature 1-2 correlation)

По данным корреляции видим, что у данных выборок значение С нерепрезентативно, значение стандартного отклонения также недостаточно большое, чтобы считать наблюдения информативными по температурным изменениям (в 12 раз ниже чем в Стяжках и лишь в 3 раза больше чем при испытаниях газов). Метки таргетов также все нулевые. По итогу, данные выборки являются малополезными и могут лишь привести к ухудшению результата.

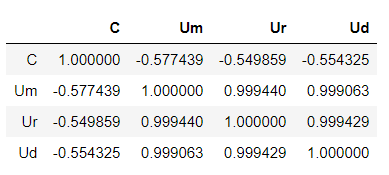
## Тест на открытом воздухе всю ночь

Рассмотрим стандартные выборочные характеристики выборок:



На открытом воздухе всю ночь (all night describe )

Рассмотрим значения корреляции признаков:



Корреляция признаков теста на открытом воздухе всю ночь (all night correlation)

Некоторое затруднение для использования данных этого наблюдения представляет значения корреляций признаков к значению С: если в газовых тестах значение корреляции было близко к 1, в других выборках было близко к нулю, в данных же наблюдениях принимает значение в районе **0.55.** Ввиду этого, а также значения таргета близкого к нулю, будем использовать лишь небольшую часть из данных наблюдений. Главная ценность данных наблюдений – это не произвольное значение С для оценки точности при близком к нулю таргете.

# Разделение данных на обучающую и тестовую выборки

Перед построением моделей разделим все наши данные на обучающую и тестовую выборки.

Первая пара обучающей и тренировочной выборок – выборки из Стяжек 1 и 2. Особенность тестовой выборки из этого набора данных в том, что она не обладает значением С, из-за чего мы не можем посчитать значение улучшения результата, а можем посчитать лишь ошибку по отношению к таргету нашей предсказательной модели.

Вторая пара – выборки из тестов с газами 1 и 2. Так как в первом тесте значения С шумовые, то весь тестовый семпл был взят из 2 теста.

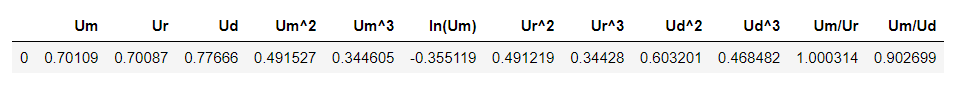
Третья пара – выборки из измерений всю ночь. Так как исходная выборка слишком большая с около нулевыми значениями таргета, а также из-за не высокой корреляции признаков со значением С, выберем небольшой семпл случайных объектов и уже его разделим на обучающую и тестовую выборки.

# Построение и испытание нейронных сетей

Для построения нейронных сетей будем использовать Keras – API над TensorFlow для эффективного построения нейронных сетей.

Перед обучением нейронной сети (далее NN), необходимо решить проблему доминирования меток нулевого таргета в нашей обучающей выборке. Для этого, из данных температурных испытаний (Стяжки) выберем случайный семпл меньшего размера, а количество объектов с ненулевым таргетом увеличим с помощью случайного выбора с возвращением.

Далее сгенерируем дополнительные признаки для нашей выборки – будем использовать квадратическую и кубическую степень наших признаков, логарифм Um, а также отношение Um к Ur и к Ud.



Нелинейные новые признаки, полученные из исходных

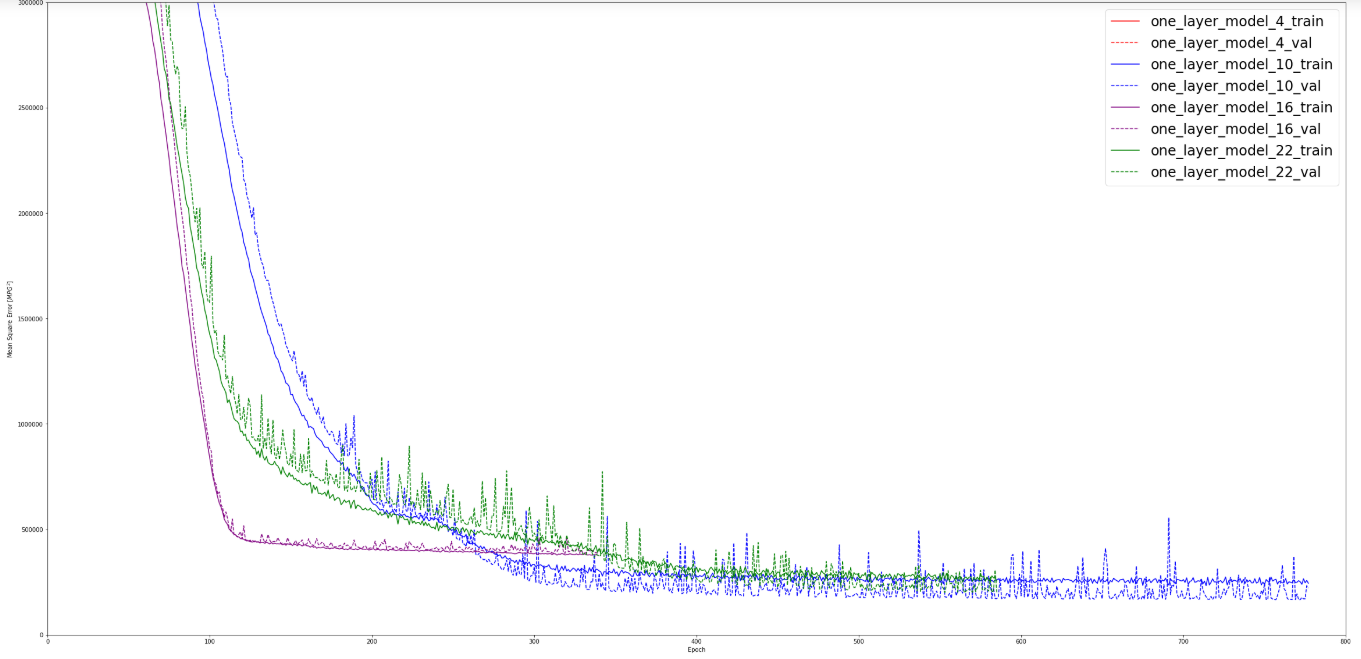
После этого необходимо разделить наши данные на обучающую и валидационную выборки и применить к ним процедуру нормализации, так как у наших новых нелинейных признаков могут быть существенно разные размерности, что негативно сказывается на обучении.

Рассмотрим работу однослойных и двуслойных сетей, с различным количеством нейронов на слое. Использование большего числа слоев не представляется целесообразным, так как сгенерировав новые признаки мы и без того перешли в достаточно сложное признаковое пространство, а NN с количеством слоев более двух рекомендуется использовать только при описании особо сложных зависимостей.

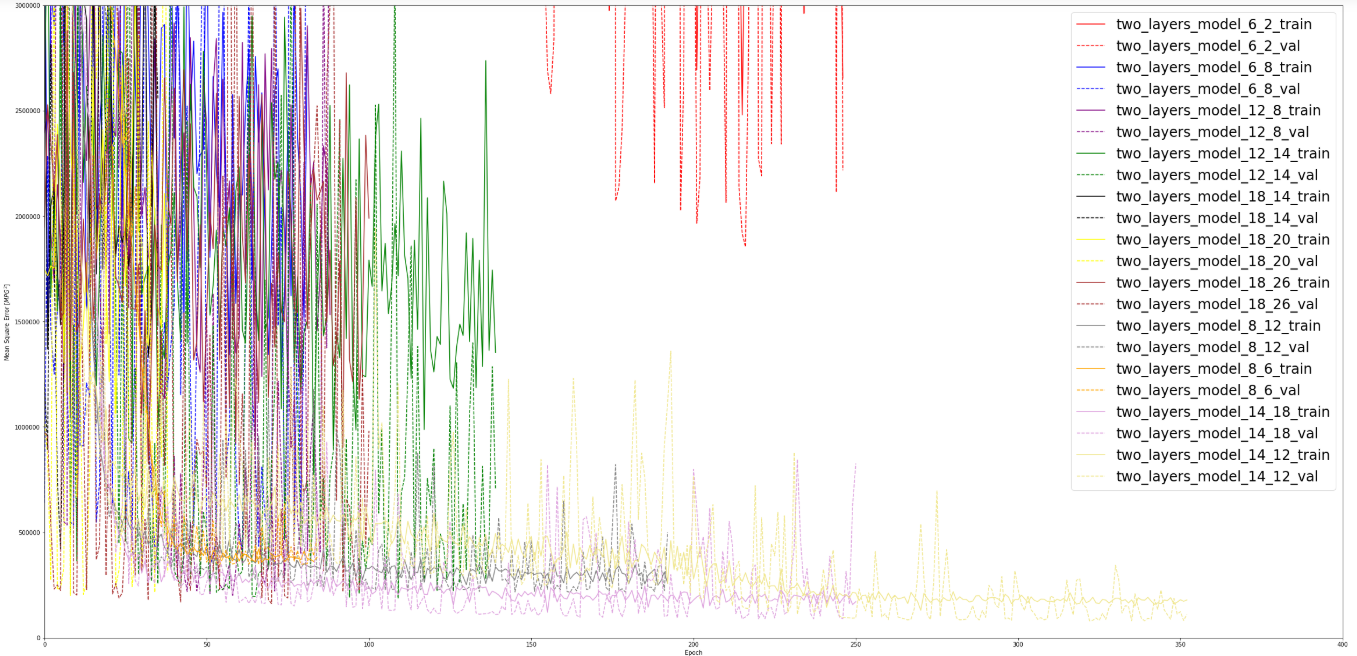
В качестве оптимизатора градиентного спуска будем использовать **RMSprop** с параметром шага **0.05.** В качестве функции активации на каждом слое (кроме выходного) будем использовать стандартный **relu**, также на каждом слое будем использовать **L1** регуляризацию с коэффициентом регуляризации **0.001**, метрика качества **MSE**.

Обучение будем проводить в размере **2000** эпох с параметром **batch size** равным **200**, с возможностью прекращения обучения, если качество прогноза на валидационной выборке не растёт на протяжении 20 эпох.

Рассмотрим графики MSE по эпохам для каждой построенной NN:



Однослойные модели



Двуслойные модели

На графиках мы видим, что все однослойные сети сходятся к одной асимптоте, со скоростью, зависящей от их сложности. С двуслойными моделями ситуация хуже – очень многие модели (особенно с небольшим числом нейронов) и вовсе не сходятся, прекращаю свою работу, не выходя на тенденцию улучшения качества. Однако определённые модели показывают очень хорошую точность.

Сравним итоговые результаты 3-х самых лучших моделей: 2-х однослойных и 1-й двуслойной.

Среди двуслойных моделей лучше всех себя показала модель с конфигурацией **14** нейронов на первом слое и **12** нейронов на втором.



Завершающие эпохи обучения двухслойной сети

Среди однослойных лучше всего себя показали 22-х и 10-ти нейронная сети:



Завершающие эпохи обучения однослойной сети 22

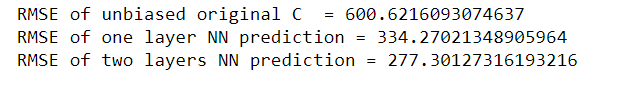


Завершающие эпохи обучения однослойной сети 10

Из таблиц видно, что минимальной ошибки достигает именно двухслойная нейронная сеть, а среди однослойных, более сложная 22 нейронная показывает себя точнее и стабильнее.

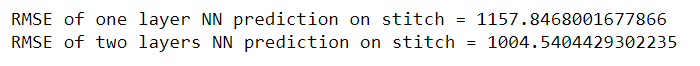
# Итоговая проверка моделей

Рассмотрим результаты, которые наши модели покажут на тестовой выборке с газами и сравним с результатом исходного показания датчика по параметру С, но откорректированного на смещение ошибки при нулевом таргете.



Как мы видим обе наши модели на тестовой выборке с газами демонстрируют точность значительно выше исходных показателей, даже откорректированных на смещение.

Теперь протестируем наши модели на температурных тестах. К сожалению, показаний датчика для сравнения у нас нет.



Мы видим, что на температурных данных ошибка достаточно высокая, но двухслойная модель всё равно показывает себя лучше, поэтому как финальную модель мы выберем именно её.

# Заключение по результатам

Построенная двухслойная нейронная сеть в указанной конфигурации **14-12** позволяет добиться значительно более высокой точности при испытаниях на газовых смесях, но по-прежнему сильно подвержена температурному влиянию. Главным барьером в достижении высокой точности на температурных тестах является малое количество данных и отсутствие разнообразия в метках целевых значений.

Для повышения качества необходимо, чтобы температурные испытания проходили при различных газовых смесях. Доминирование нулевых меток в выборках лишь мешает моделям выявлять глобальные закономерности. Также желательно увеличить верхний предел значения доли метана в газовых испытаниях, так как малая вариативность таргетов, и низкий верхний предел значений, сильно ограничивают в возможных применимых подходах к задаче регрессии, делая недоступными для использования широкий класс моделей на основе решающих деревьев.

На данный же момент оптимальным вариантом являлась бы взвешенная сумма откорректированных на смещение ошибки показаний датчика и показаний предоставляемых моделью, с возможным сглаживанием случайных шумов с помощью скользящего среднего с небольшим окном.

# Список литературы

1. sklearn documentation - <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>
2. The Number of Neurons in the Hidden Layers - <https://www.heatonresearch.com/2017/06/01/hidden-layers.html>
3. TensorFlow Guide in using Keras for solving regression tasks - <https://www.tensorflow.org/tutorials/keras/regression?hl=ru>
4. Implementing a Neural Network for Regression - <https://www.pyimagesearch.com/2019/01/21/regression-with-keras/>
5. A New Approach to Linear Filtering and Prediction Problems - <http://www.cs.unc.edu/~welch/kalman/media/pdf/Kalman1960.pdf>
6. Курс МФТИ по машинному обучению - <https://www.youtube.com/watch?v=JGIOoMRFPKk&list=PL4_hYwCyhAvasRqzz4w562ce0esEwS0Mt>
7. Код работы на github - <https://github.com/dkamianskii/LEDSensorsMeasuringAnalysis>