МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет

имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Прогнозирование конечных свойств

новых материалов (композиционных материалов)

Слушатель Слепнёв Петр Юрьевич

Москва, 2022

Оглавление

[**1.** **Введение** 3](#_Toc106198621)

[**2.** **Аналитическая часть** 5](#_Toc106198622)

[2.1. Теоретическая часть решаемой задачи 5](#_Toc106198623)

[2.2. Предварительный анализ исходных данных 7](#_Toc106198624)

[2.3. Разведочный анализ данных 13](#_Toc106198625)

[**3.** **Практическая часть** 20](#_Toc106198626)

[3.1. Постановка задачи 20](#_Toc106198627)

[3.2. Модели, предполагаемые к использованию для решения задач 26](#_Toc106198628)

[3.3. Подготовка данных 29](#_Toc106198629)

[3.4. Моделирование 30](#_Toc106198630)

1. **Введение**

***Общая часть***

В ходе данной работы слушателем были изучены свойства композитных материалов, проведена работа по анализу полученных в ходе опытов характеристик полученных материалов.

Композиционный материал – конструкционный (металлический или неметаллический) материал, в котором имеются усиливающие его элементы в виде нитей, волокон или хлопьев более прочного материала. Примеры композиционных материалов: пластик, армированный борными, углеродными, стеклянными волокнами, жгутами или тканями на их основе; алюминий, армированный нитями стали, бериллия. Комбинируя объемное содержание компонентов, можно получать композиционные материалы с требуемыми значениями прочности, жаропрочности, модуля упругости, абразивной стойкости, а также создавать композиции с необходимыми магнитными, диэлектрическими, радиопоглощающими и другими специальными свойствами.

***Типы композиционных материалов***

Композиционные материалы с металлической матрицей.

Композитные материалы или композиционные материалы состоят из металлической матрицы (чаще Al, Mg, Ni и их сплавы), упрочненной высокопрочными волокнами (волокнистые материалы) или тонкодисперсными тугоплавкими частицами, не растворяющимися в основном металле (дисперсно-упрочненные материалы). Металлическая матрица связывает волокна (дисперсные частицы) в единое целое. Волокно (дисперсные частицы) плюс связка (матрица), составляющие ту или иную композицию, получили название композиционные материалы.

Композиционные материалы с неметаллической матрицей.

Композиционные материалы с неметаллической матрицей нашли широкое применение. В качестве неметаллических матриц используют полимерные, углеродные и керамические материалы. Из полимерных матриц наибольшее распространение получили эпоксидная, фенолоформальдегидная и полиамидная.

Угольные матрицы коксованные или пироуглеродные получают из синтетических полимеров, подвергнутых пиролизу. Матрица связывает композицию, придавая ей форму. Упрочнителями служат волокна: стеклянные, углеродные, борные, органические, на основе нитевидных кристаллов (оксидов, карбидов, боридов, нитридов и других), а также металлические (проволоки), обладающие высокой прочностью и жесткостью.

Свойства композиционных материалов зависят от состава компонентов, их сочетания, количественного соотношения и прочности связи между ними.

Армирующие материалы могут быть в виде волокон, жгутов, нитей, лент, многослойных тканей.

***Цели работы***

При проектировании новых композитных материалов основной задачей является определение их свойств с наименьшими трудовыми и финансовыми затратами.

Указанная задача может быть решена путем проведения физических экспериментов с целью определения свойств материалов или построение математических прогнозных моделей.

В рамках данной работы мы предпримем попытку анализа модуля упругости при растяжении и прочности при растяжении, а также попробуем написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.

1. **Аналитическая часть**

## Теоретическая часть решаемой задачи

Предполагается, что в данной задаче речь идет о запатентованной специалистами МИЦ «Композиты России» МГТУ им. Н. Э. Баумана способе изготовления волокнистых заготовок формы, относящийся к технологии направленной укладки волокон (Tailored Fiber Placement — TFP).

При использовании данного способа повышение физико-механических и эксплуатационных характеристик изделий происходит в результате оптимизации технологических процессов нашивки и пропитки.

Патент был получен в ходе масштабной разработки технологии изготовления объемно-армированных преформ элементов газотурбинных двигателей (ГТД) из полимерных композиционных материалов (ПКМ), которая представляет собой автоматизированную направленную нашивку углеродного ровинга на водорастворимую подложку на основе поливинилового спирта (PVA). Помимо достижения оптимальных для эксплуатации совокупности физико-механических показателей, технология имеет такие преимущества, как возможность изготовления серий изделий с идентичными геометрическими, весовыми и прочностными характеристиками, отсутствие человеческого фактора, экономичность и безотходность производства.

Оптимизация уже запатентованного процесса происходит за счет применения рациональных параметров операции нашивки преформ изделий из ПКМ: плотности укладки армирующих углеродных волокон и длины прошивного зигзагообразного стежка, которые обеспечивают наибольшую скорость пропитки структурного каркаса и качество образовавшегося композита.

В предлагаемом способе нашивка волокон осуществляется на подложку из водорастворимого материала на основе PVA, которая впоследствии легко отделяется от заготовки вымыванием, не нарушая её целостности и не вызывая разрушение волокон.

Технология TFP представляет собой нашивку нитеообразного материала на основание с целью получения заготовки. (рис. 1)

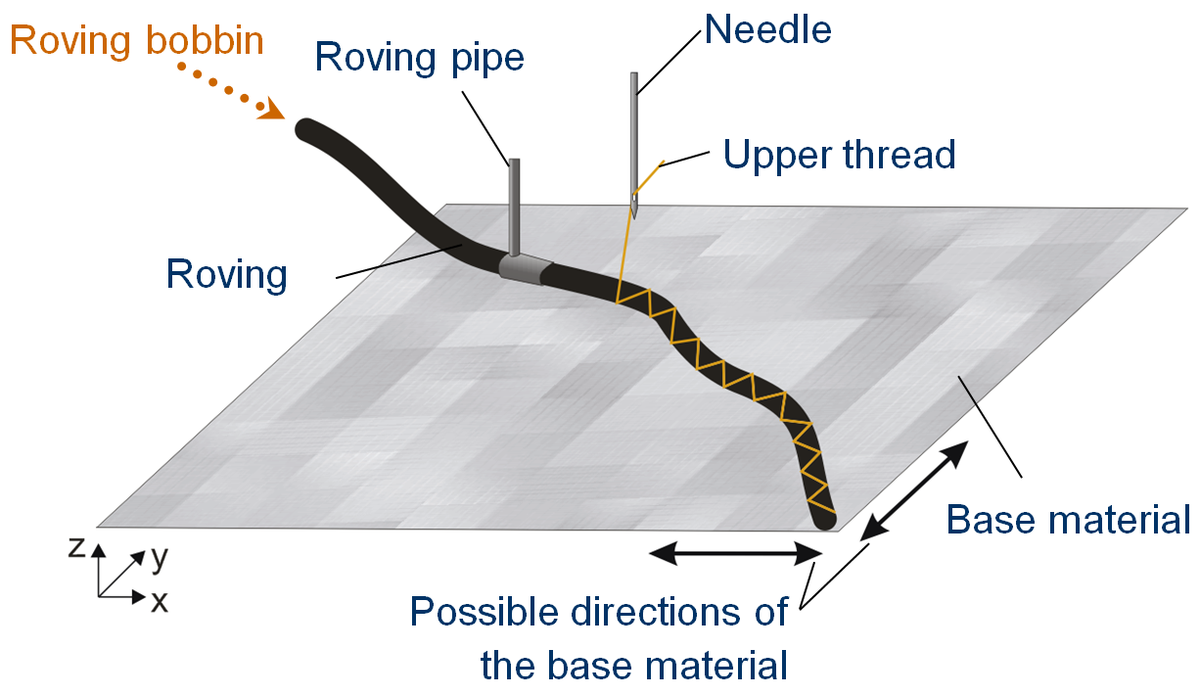


Рис.1

При этом, «вышитая» деталь может быть абсолютно любой формы в том числе, в 3-d формате. Пример нашивки – рис.2

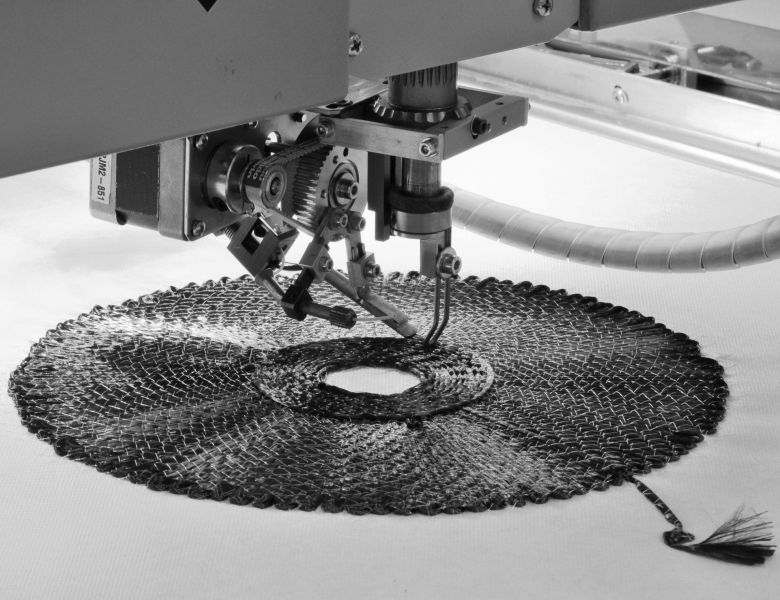


Рис.2

## Предварительный анализ исходных данных

Предварительный анализ файла X\_bp.xlsx.

После объединения данных появилось предположение, что датасет в части данных файла «X\_bp.xlsx» с 23 строки по 1022 был расширен. Данный вывод был сделан на основании следующих предположений:

* Визуальный анализ показывает, что данные с 23 по 1022 строки имеют другую структуру (9 точек после запятой)
* Выделенные данные (рис. 3) имеют одинаковые значения и одинаковую структуру (9 точек после запятой). Можно сделать предположение, что эти данные были добавлены с целью заполнения пустот. Одним из способов заполнения пустот является заполнение средними значениями. Проверка средних значений показала, что они были рассчитаны не на основании всего датасета (1022 строки), а на основании лишь первых 23 строк.



Рис.3

* Был произведен анализ файла по рекомендованной в ходе обучения методике (Рис.4)

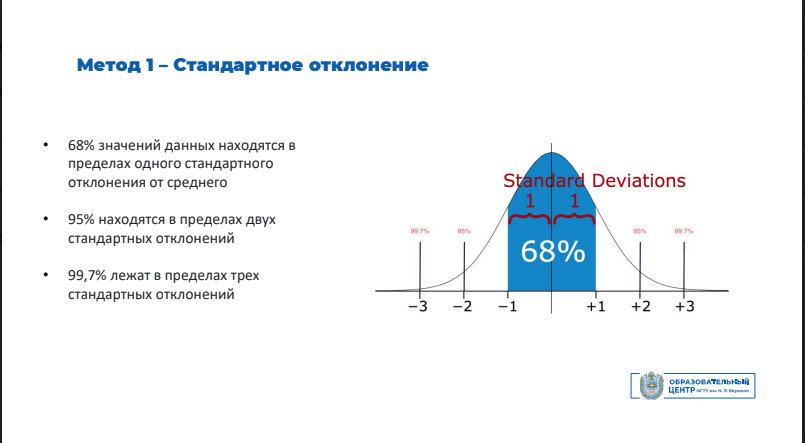


Рис.4

Было выяснено, что данные практически точно принадлежат рекомендуемым промежуткам, что может косвенно говорить о том, что все данные были сгенерированы автоматически с соблюдением нормального распределения (рис.5).



Рис. 5

* В связи с тем, что в работе рассматриваются материалы с определенными физическими свойствами, то генерация новых данных должна быть осуществлена с соблюдением физических характеристик рассматриваемых материалов. В то же время отношение минимального значения к максимальному составляет по показателям от 1,3 до 2318 раз (рис.6), что ставит под сомнение физическую сущность эксперимента.



Рис.6

Корректировка файла X\_bp.xlsx

Проведем предварительную корректировку данных по первым 23 строкам, которые представляются полученными в результате реальных экспериментов

Столбец «Температура вспышки С\_2». Из 23 значений только одно значение, отличное от «300» по строке 0. Логично предположить, что это значение либо выброс, либо ошибка. В случае удаления этого значения, все остальные значения становятся равными 3000С. Скорее всего, опыты проводились при одинаковой температуре 3000С. Таким образом, более вероятно, что пропущенные значения и выбросы будут равны 3000С.

Столбцы «Поверхностная плотность, г/м2», «Модуль упругости при растяжении, Гпа», «Прочность при растяжении, Мпа», «Потребление смолы, г/м2», по-видимому, отражают характеристики базальтовой ткани.

а) Ткань 1 (Рис.7)

Поверхностная плотность, г/м2 - 210

Модуль упругости при растяжении, Гпа - 70

Прочность при растяжении, Мпа - 3000

Потребление смолы, г/м2 – 220



Рис.7

б) Ткань 2 (рис. 8)

Поверхностная плотность, г/м2 - 380

Модуль упругости при растяжении, Гпа - 75

Прочность при растяжении, Мпа - 1800

Потребление смолы, г/м2 – 120



Рис.8

в) Ткань 3 (Рис.9)

Поверхностная плотность, г/м2 - 470

Модуль упругости при растяжении, Гпа – 73,333333333 (рассчитано как среднее, оставим в таком виде, т.к. это самый оптимальный способ заполнения указанных данных)

Прочность при растяжении, Мпа – 2455,5555556 (рассчитано как среднее, оставим в таком виде, т.к. это самый оптимальный способ заполнения указанных данных)

Потребление смолы, г/м2 - 220



Рис.9

г) Ткань 4 (Рис.10)

Поверхностная плотность, г/м2 - 1010

Модуль упругости при растяжении, Гпа – 78

Прочность при растяжении, Мпа – 2000

Потребление смолы, г/м2 - 300



Рис.10

Столбцы «Количество отвердителя, м.%», «Содержание эпоксидных групп 2» нуждаются в точечной корректировке. Исходя из сути опыта маловероятно, что эпоксидной смолой пропитывали при разных условиях. Скорее всего, меняли нашивку при одинаковых условиях пропитки смолой. Поэтому, предполагаю, что нужно оставить одинаковые значения в этих столбцах.

Откорректированная таблица значений 0-22 выглядит следующим образом.



Рис.11

**Предварительный анализ файла X\_nup.xlsx.**

Расчетные значения начинаются с 40 значения

Значения 0-18 абсолютно такие же как значения, только отличаются от 20 – 39. Отличия только в показателе «угол нашивки»

Значение 19 – абсолютно нулевое, но т.к. оно есть по этому ключу присоединяется соответствующее значение из таблицы X\_bp.xlsx, то оно должно быть заполнено данными. Можно заметить, что значение с шагом нашивки, равное 4 имеет только 3 плотности нашивки, тогда как все остальные – 4. Можно предположить, что значение 19 является выбросом, и его необходимо поставить под номером 0 и присвоить ему шаг нашивки 4 и плотность 47. Нумерацию сдвигать не будем.



Рис.12

## Разведочный анализ данных

Посмотрим нулевые значения в объединенном датасете

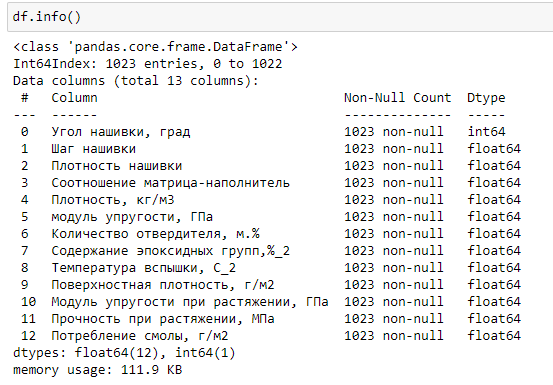


Рис.13

Нулевые значения отсутствуют.

Посмотрим выбросы в данных.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Рис.14

По данным можно заметить, что существуют выбросы данных, которые необходимо удалить.

Посмотрим, насколько нормально распределены данные.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Рис.15

По представленных данным видно, что значения имеют, в основном, нормальное распределение.

Сформируем попарный график рассеяния всех показателей.

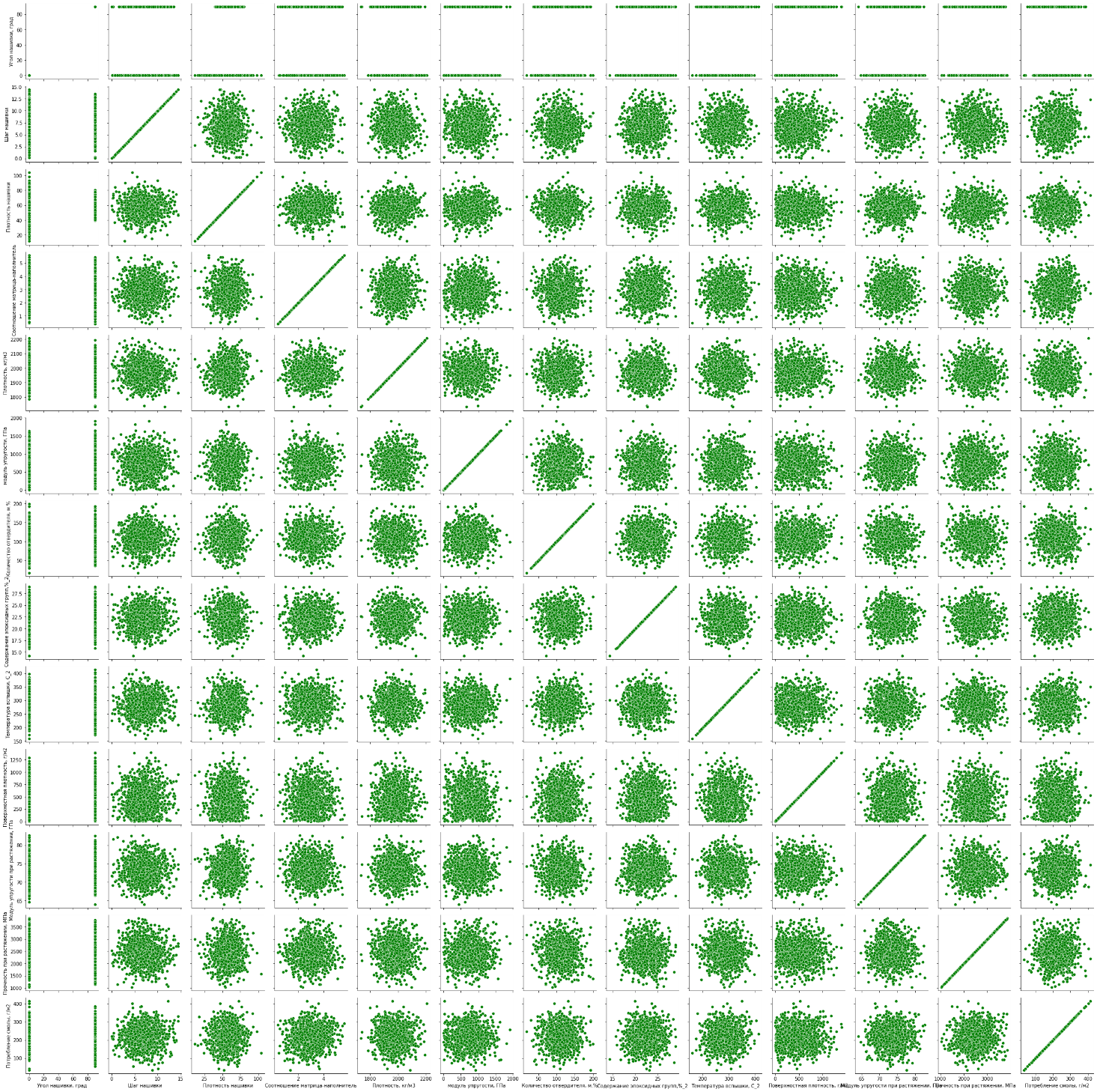


Рис 16.

По графику не прослеживается зависимости показателей друг от друга

Построим тепловую карту корреляции показателей

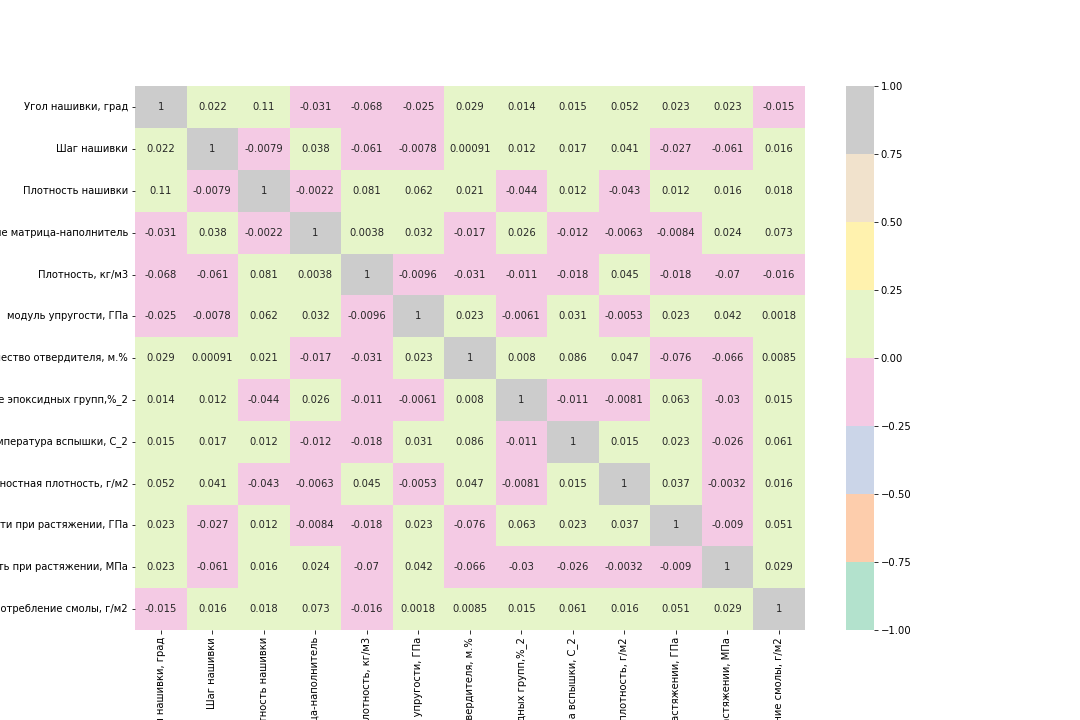


Рис 17

По графику видно, что попарная корреляция близка к нулю. Показатели не коррелируют друг с другом

По разведочному анализу можно сделать вывод, что, несмотря на нормальное распределение данных, прогнозирование по этим данным представляется затруднительным.

1. **Практическая часть**

## Постановка задачи

В связи с тем, что есть предположение о то, что данные, начиная с 23 значения, являются расчетными, предлагается рассмотреть несколько моделей. При этом, в случае неудачи с исходной моделью, ответить на вопрос, при каком уровне изменения данных может быть получен приемлемый результат.

В связи с этим предполагается подготовить 4 варианта исходных данных

1. Исходный предложенный вариант. Подразумевает полностью исходные данные, из которых исключены выбросы. Формула исключения выбросов

max = q75+(1,5\*(q75-q25))

min = q25-(1,5\*(q75-q25))

1. Вариант №1, который будет подготовлен следующим образом:

* Формируем 23 таблицы, предполагая, что первые 23 строки – эталонные
* По методу среднеквадратичного отклонения заполняем эти таблицы 44 строками (11 последних отбросим) таким образом, чтобы строка №2 была максимально похожа на эталонную, №3 - менее и так далее.
* В каждой таблице по первым **20** строкам поставим равным 0 значения по столбцам, выходящие за пределы

max = q75+(1\*(q75-q25))

min = q25-(1\*(q75-q25))

Больше значения равного 1 ставить нет смысла, т.к. выборка значений - небольшая

* Заменим нули средними значениями по столбцу (по всем оставшимся строкам)
* Посчитаем процент замены

1. Вариант №2. Аналогичен варианту №1. Порядок замены:

* В каждой таблице по первым **10** строкам поставим равным 0 значения по столбцам, выходящие за пределы

max = q75+(**0,5**\*(q75-q25))

min = q25-(**0,5**\*(q75-q25))

1. Вариант №3 – самый радикальный. Порядок замены:

* В каждой таблице по первым **5** строкам поставим равным 0 значения по столбцам, выходящие за пределы

max = q75+(**0**\*(q75-q25))

min = q25-(**0**\*(q75-q25))

Также проверим гипотезы «физической» природы данных.

1. Предположим, что все эксперименты осуществлялись при одних и тех же параметрах показателей «**Количество отвердителя, м.%»,** «**Содержание эпоксидных групп, %\_2», «Температура вспышки, С\_2»,** соответственно, эти параметры можно исключить.
2. Предположим, что, исходя из сути эксперимента, показатели **«Модуль упругости при растяжении, Гпа»** и **«Прочность при растяжении, Мпа»** являются характеристиками нити, из которой делается нашивка, эти данные могут не зависеть от

Соотношение матрица-наполнитель

Плотность, кг/м3

модуль упругости, ГПа

Количество отвердителя, м.%

Содержание эпоксидных групп, %\_2

Температура вспышки, С\_2

Обе эти «физические» гипотезы будем проверять на основном наборе данных.

Итого, у нас получается 17 моделей, которых мы хотим проверить, которые можно свести в таблицу

Таблица 1

| **№ п/п** | **X** | **Y** | **Набор данных** | **Аббревиатура в коде** |
| --- | --- | --- | --- | --- |
| 1 | Все остальное | 'Модуль упругости при растяжении, ГПа' | Исходный | M11 |
| 2 | Все остальное | 'Модуль упругости при растяжении, ГПа' | Вариант 1 | M11\_2 |
| 3 | Все остальное | 'Модуль упругости при растяжении, ГПа' | Вариант 2 | M11\_3 |
| 4 | Все остальное | 'Модуль упругости при растяжении, ГПа' | Вариант 3 | M11\_4 |
| 5 | Все остальное | Прочность при растяжении, МПа | Исходный | M21 |
| 6 | Все остальное | Прочность при растяжении, МПа' | Вариант 1 | M21\_2 |
| 7 | Все остальное | Прочность при растяжении, МПа | Вариант 2 | M21\_3 |
| 8 | Все остальное | Прочность при растяжении, МПа' | Вариант 3 | M21\_4 |
| 9 | Все остальное | Соотношение матрица-наполнитель | Исходный | M31 |
| 10 | Все остальное | Соотношение матрица-наполнитель' | Вариант 1 | M31\_2 |
| 11 | Все остальное | Соотношение матрица-наполнитель | Вариант 2 | M31\_3 |
| 12 | Все остальное | Соотношение матрица-наполнитель | Вариант 3 | M31\_4 |
| 13 | Все, за вычетом «Количество отвердителя, м.%», «Содержание эпоксидных групп, %\_2», «Температура вспышки, С\_2» | 'Модуль упругости при растяжении, ГПа' | Исходный | M21 |
| 14 | Все, за вычетом «Количество отвердителя, м.%», «Содержание эпоксидных групп, %\_2», «Температура вспышки, С\_2» | Прочность при растяжении, МПа' | Исходный | M22 |
| 15 | Все, за вычетом «Количество отвердителя, м.%», «Содержание эпоксидных групп, %\_2», «Температура вспышки, С\_2» | Соотношение матрица-наполнитель | Исходный | M23 |
| 16 | Все, кроме  Соотношение матрица-наполнитель  Плотность, кг/м3  модуль упругости, ГПа  Количество отвердителя, м.%  Содержание эпоксидных групп, %\_2  Температура вспышки, С\_2 | 'Модуль упругости при растяжении, ГПа' | Исходный | M31 |
| 17 | Все, кроме  Соотношение матрица-наполнитель  Плотность, кг/м3  модуль упругости, ГПа  Количество отвердителя, м.%  Содержание эпоксидных групп, %\_2  Температура вспышки, С\_2 | Прочность при растяжении, МПа' | Исходный | M32 |

Итогом решения этих задач планируется следующее:

* Проверка гипотез о физической природе некоторых данных
* Ранжирование показателей по степени приемлемости в зависимости от процента изменения исходного датасета. Для данного исследования планируется 3 изменения датасета.

## Модели, предполагаемые к использованию для решения задач

Мы будем использовать следующие методы:

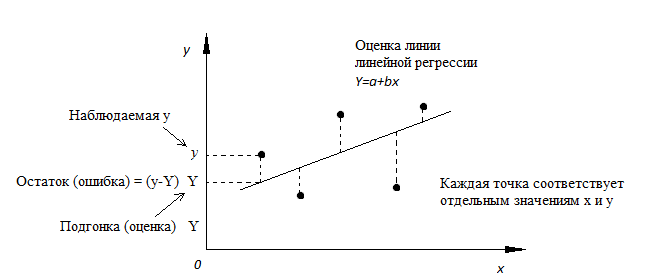
**LinearRegression()** – простая линейная регрессия.

Мы выполняем регрессионный анализ, используя выборку наблюдений, где a и b – выборочные оценки истинных (генеральных) параметров, α и β, которые определяют линию линейной регрессии в популяции (генеральной совокупности).

Наиболее простым методом определения коэффициентов a и b является метод наименьших квадратов (МНК).

Подгонка оценивается, рассматривая остатки (вертикальное расстояние каждой точки от линии, например, остаток = наблюдаемый y – предсказанный y, см рис.

Линию лучшей подгонки выбирают так, чтобы сумма квадратов остатков была минимальной.

Рис. 18

**Ridge() -** Гребневая регрессия или ридж-регрессия - один из методов понижения размерности. Применяется для борьбы с избыточностью данных, когда независимые переменные коррелируют друг с другом, вследствие чего проявляется неустойчивость оценок коэффициентов многомерной линейной регрессии. Возможно, при корректировке данных может появиться такая зависимость признаков, в связи с этим предполагается использовать данный метод.

Метод **Lasso()** - метод регрессии лассо (LASSO, Least Absolute Shrinkage and Selection Operator) — это вариация линейной регрессии, специально адаптированная для данных, которые демонстрируют сильную мультиколлинеарность (то есть сильную корреляцию признаков друг с другом).

Метод регрессии лассо (LASSO, Least Absolute Shrinkage and Selection Operator) — это вариация линейной регрессии, специально адаптированная для данных, которые демонстрируют сильную мультиколлинеарность (то есть сильную корреляцию признаков друг с другом).

Метод пытается уменьшить сложность данных так, чтобы их можно было обрабатывать простыми методами регрессии, искривляя пространство, на котором они лежат. В этом процессе лассо автоматически помогает устранить или исказить сильно коррелированные и избыточные функции в методе с низкой дисперсией.

Регрессия лассо использует регуляризацию L1, то есть взвешивает ошибки по их абсолютному значению. Вместо, например, регуляризации L2, которая взвешивает ошибки по их квадрату, чтобы сильнее наказывать за более значительные ошибки.

Такая регуляризация часто приводит к более разреженным моделям с меньшим количеством коэффициентов, так как некоторые коэффициенты могут стать нулевыми и, следовательно, будут исключены из модели. Это позволяет ее интерпретировать.

Метод **SVR() –** Метод опорных векторов. В основе метода опорных векторов для задач регрессии или регрессии опорных векторов (SVR) лежит поиск гиперплоскости, при которой риск в многомерном пространстве будет минимальным. По сравнению с традиционной регрессионной моделью SVR оценивает коэффициенты путем минимизации квадратичных потерь. Так, если прогнозное значение попадает в область гиперплоскости, то потери равны нулю. В противном случае разности прогнозного и фактического значений.

Метод **DecisionTreeRegressor (дерево решений)** - Для каждого атрибута в наборе данных алгоритм формирует узел, где наиболее важный атрибут помещается в корневой узел. Для оценки мы начинаем с корневого узла и продвигаемся вниз по дереву, следуя за соответствующим узлом, который удовлетворяет нашему условию или “решению”. Этот процесс продолжается до тех пор, пока не будет достигнут конечный узел, содержащий предсказание или результат дерева решений. Для оценки эффективности регрессионного алгоритма обычно используются следующие метрики: средняя абсолютная ошибка, среднеквадратичная ошибка.

Метод **RandomForestRegressor (Случайный лес)** — модель, состоящая из множества деревьев решений. Вместо того, чтобы просто усреднять прогнозы разных деревьев (такая концепция называется просто «лес»), эта модель использует две ключевые концепции, которые и делают этот лес случайным.

Случайная выборка образцов из набора данных при построении деревьев.

При разделении узлов выбираются случайные наборы параметров.

Метод **AdaBoostRegressor()** - бустинговый метод, применяемый к любой регрессионной модели (по умолчанию в данном случае к RandomForestRegressor)

Бустинг – это продвинутый алгоритм машинного обучения для решения задач классификации и регрессии. Он строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений. Из нескольких слабых моделей в итоге мы собираем одну, но уже эффективную. Общая идея алгоритма – последовательное применение предиктора (предсказателя) таким образом, что каждая последующая модель сводит ошибку предыдущей к минимуму.

К каждой модели будет осуществляться кросс-валидация в количестве 10 выборок. Это когда случайным образом выбирается какое-то количество наборов данных, по ним производится предсказания, а ошибка усредняется.

Также к каждой модели произведем усовершенствование параметров, которые выбраны случайным образом исходя из вычислительных возможностей оборудования.

## Подготовка данных

* По варианту исходных данных после удаления выбросов осталось 939 строк. Таким образом, потеря данных – 8.2%

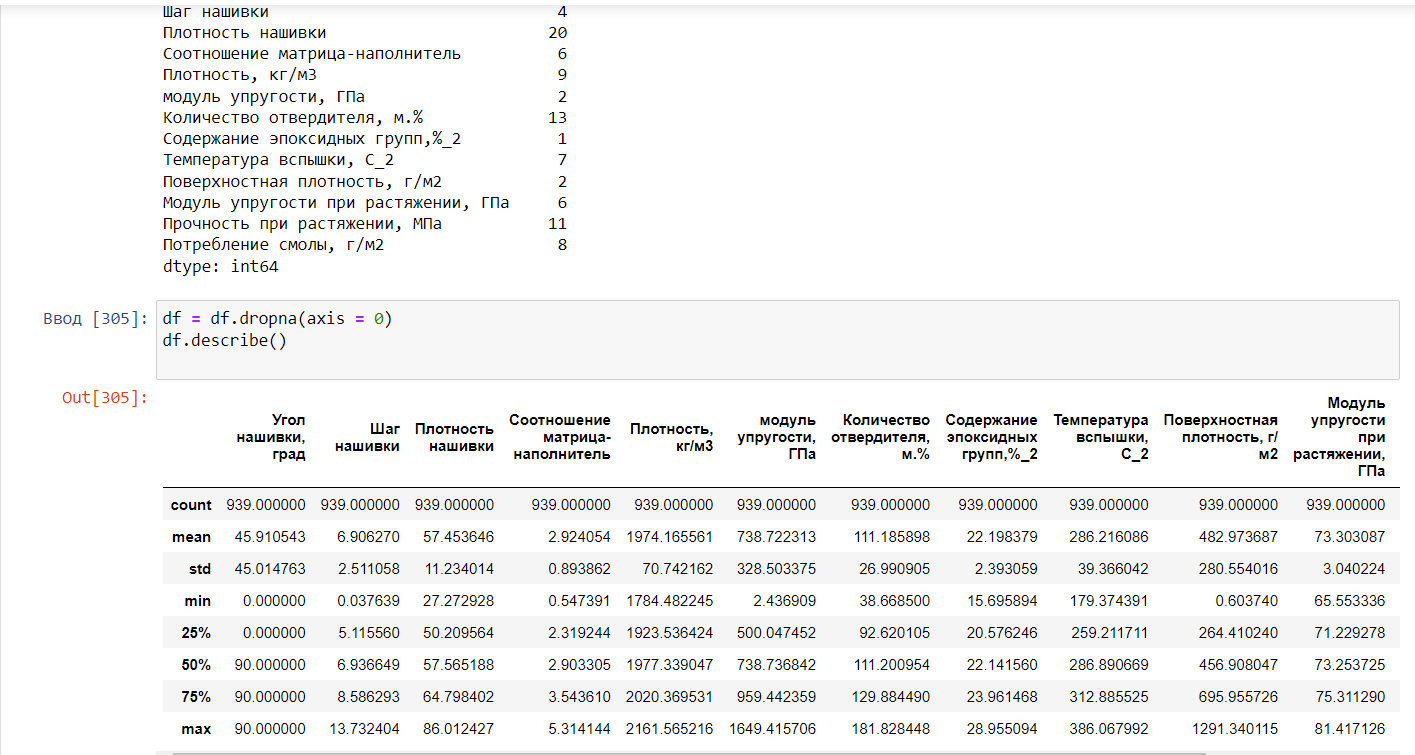
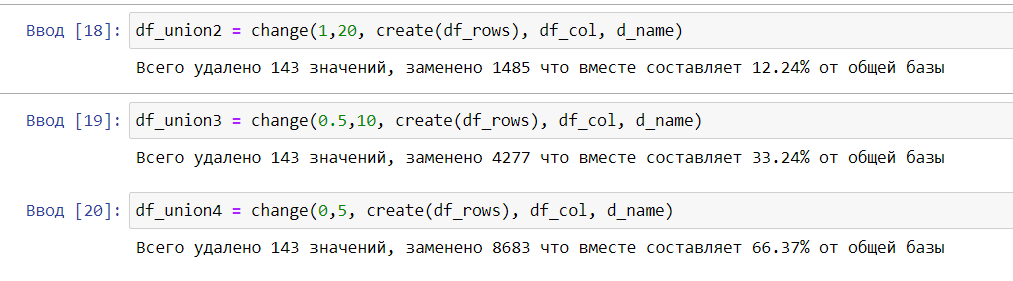


Рис.19

* По вариантам исходных данных 2,3,4 потеря составила 12,24%, 33,24% и 66,37% соответственно

Рис. 20

## Моделирование

Анализ прогнозирования показателя **«Модуль упругости при растяжении, Гпа»**

Результаты анализа встроенным набором показателей моделей показывают, что приемлемый уровень прогнозирования достигается при   
66% - ой корректировке данных. При 33% корректировке точность по коэффициенту детерминации R2 составляет 0,13, что очень низко. Наиболее точный метод – метод опорных векторов. R2 =0.73. Может считаться более-менее приемлемой точностью

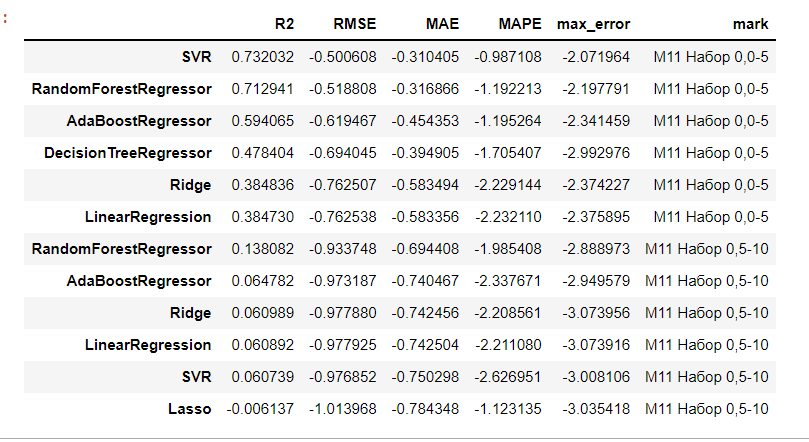
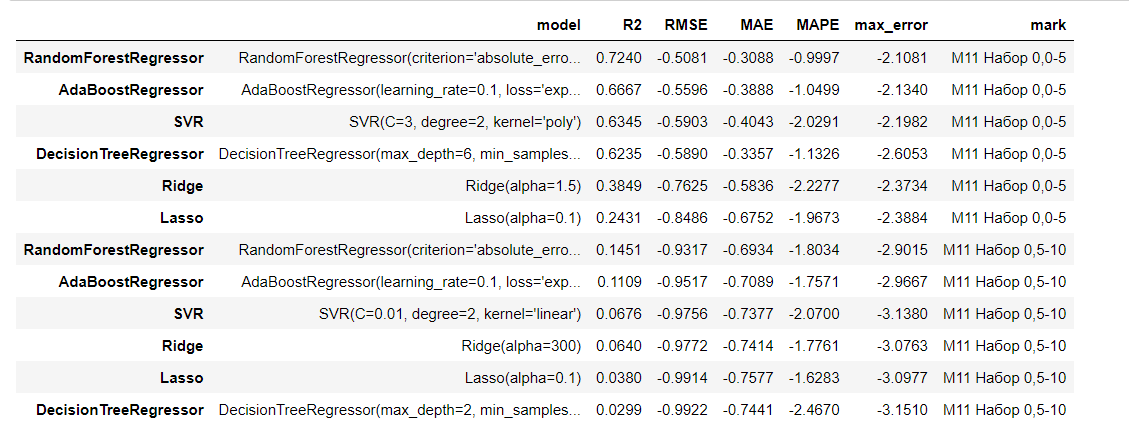


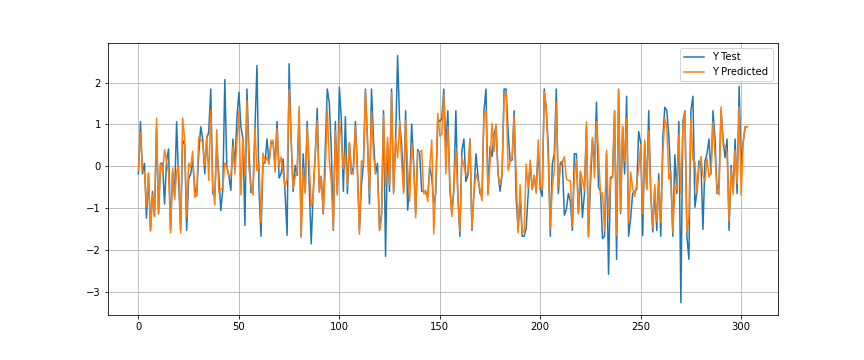
Рис. 21

Методы улучшения не дали по исходным данным с 66% потерей данных. Однако, лучшая модель с 33% потерей данных RandomForestRegressor (criterion='absolute\_error', max\_features='sqrt', n\_estimators=150).

 Рис. 22

Распечатаем график предсказания по модели

Рис. 23



Вывод: прогнозирование показателя **«Модуль упругости при растяжении, Гпа»** при текущих исходных данных не дает требуемой точности. Требуется серьезная модернизация исходных данных, либо добавление других параметров для прогнозирования. При оптимизации алгоритма корректировки данных возможно снижение процента потерь. Также возможно построение графика зависимости процента потерь от точности показателя с целью выбора оптимальных параметров для исследователя.

Анализ прогнозирования показателя **«Прочность при растяжении, МПа»**

Результаты анализа встроенным набором показателей моделей показывают, что приемлемый уровень прогнозирования достигается при   
33% - ой корректировке данных. При 12% корректировке точность по коэффициенту детерминации R2 составляет 0,18, что очень низко. Наиболее точный метод – метод опорных векторов. R2 =0.97. Такая точность может считаться хорошим результатом.

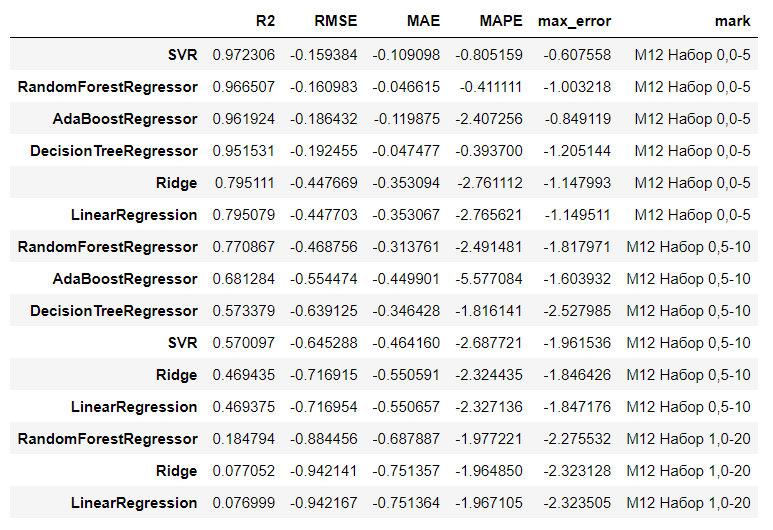


Рис. 24

Таблица улучшения моделей показывает, что точность при потере данных в 66% увеличивается за счет применения модели RandomForestRegressor(), но для данных при 33% потере лучший результат при использовании базовых моделей

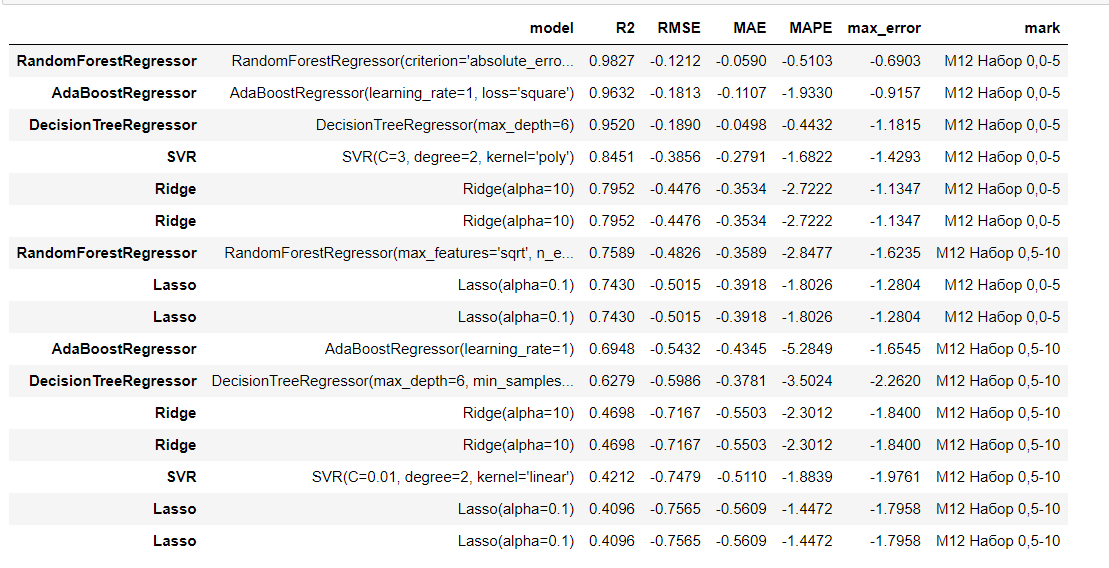
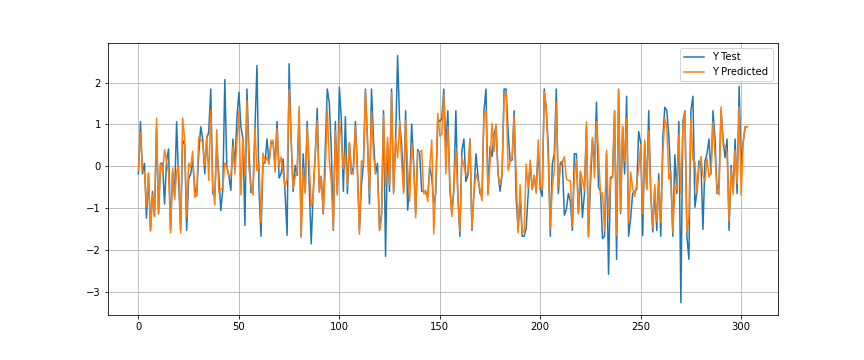


Рис. 25

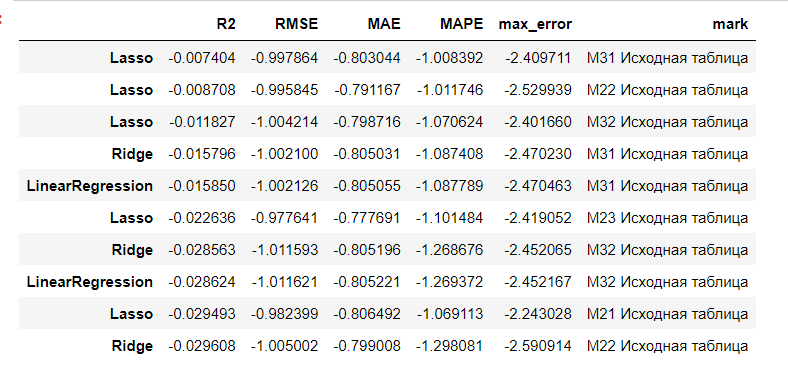
Распечатаем график предсказания по модели (33% потеря данных)

Рис. 26



Вывод по прогнозированию показателя **«Прочность при растяжении, МПа».** Показатель неплохо прогнозируется. При оптимизации алгоритма корректировки данных можно добиться неплохих результатов даже при потере данных в районе 20%

Прогнозирование показателей **«Прочность при растяжении, МПа»** и **«Модуль упругости при растяжении, Гпа»**, исходя из их физических свойств.

**** Рис. 27

**Вывод.** Гипотеза о возможности такого прогнозирования не подтвердилась. Возможно, прогнозирование на доработанных данных может дать результат, но это может быть темой отдельного углубленного исследования и, исходя из условий задачи, для такого исследования не хватает сведений. Следует заметить, что лучшие показатели lasso и Ridge, используемые для борьбы с мультиколлинеарностью. Возможно, эти показатели имеют все-таки слабую связь.

Прогнозирование показателя «**Соотношение матрица-наполнитель»**

Модель с базовыми показателями дает не очень значительную точность (R2 = 0,547) даже на исходных данных с самой большой корректировкой данных (66%), что косвенно может судить о том, что эти данные были посчитаны искусственно (они рассчитаны даже для первых 23 значений). Рекомендовать этот показатель для прогнозирования нельзя.

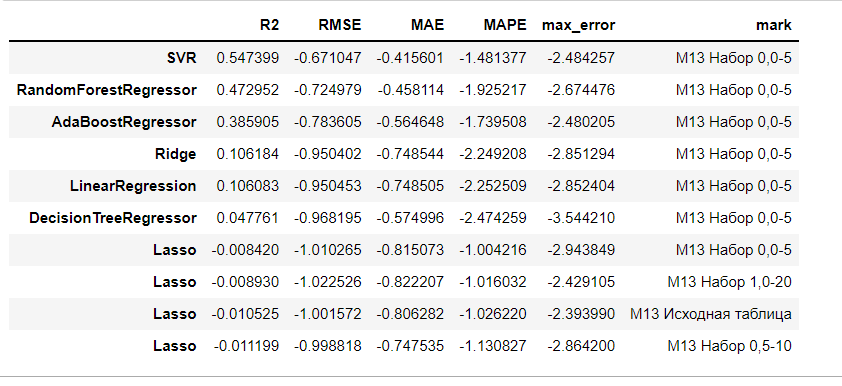
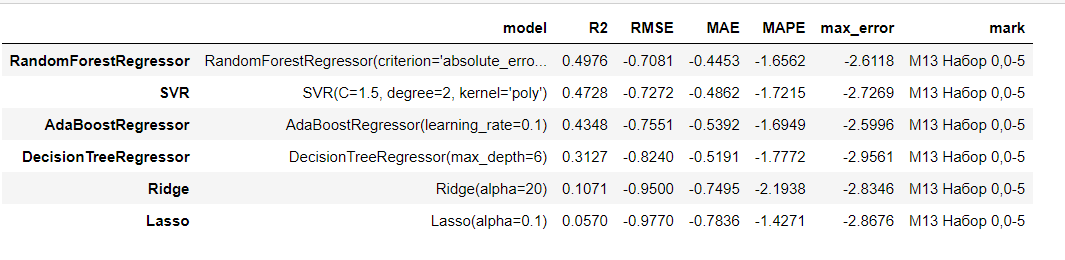


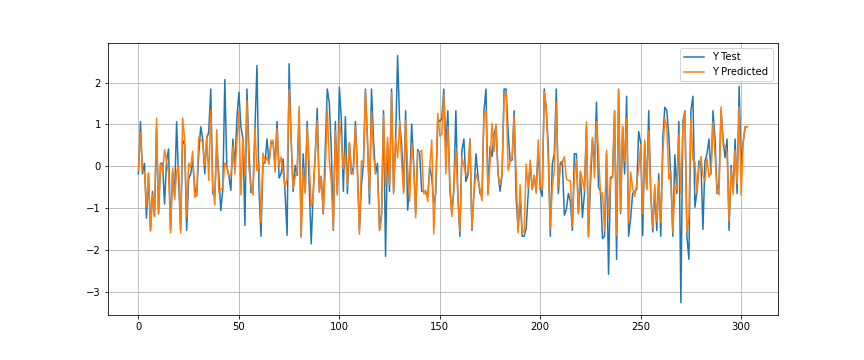
Рис. 28

Улучшение моделей для прогнозирования также не дает существенного улучшения

 Рис. 29

Ниже представлен график прогнозирования исходя из 66% потери данных базовым методом SVR()

Рис. 30



Формирование нейронной сети не улучшило ситуацию. Нейронную сеть подобрать не удалось. Были сформированы 3 нейронных сети (см блокнот), каждая с нормализующим слоем и без него. В качестве исходных данных для нейронных сетей были использованы данные с потерей 33% и 66%. Уже стандартизованные, а также данные, нормализуемые на входе в нейронную сеть. Результаты сведены в таблицу. Лучшая сеть плохо прогнозирует данные. Сети с нормализующим слоем лучше справились с задачей.

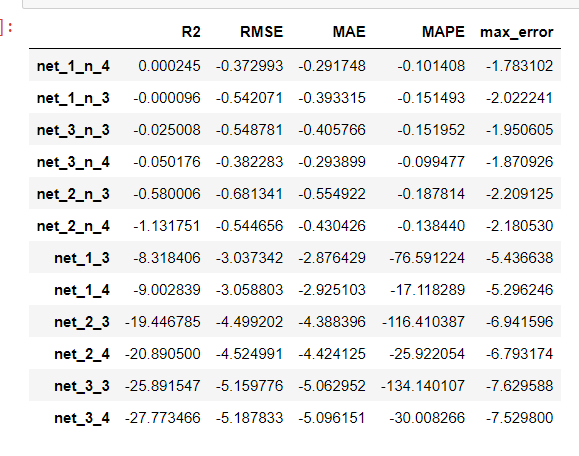


Рис. 31

График обучения сети

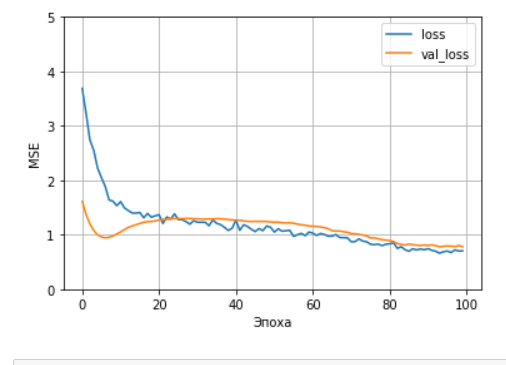


Рис. 32

График построения данных

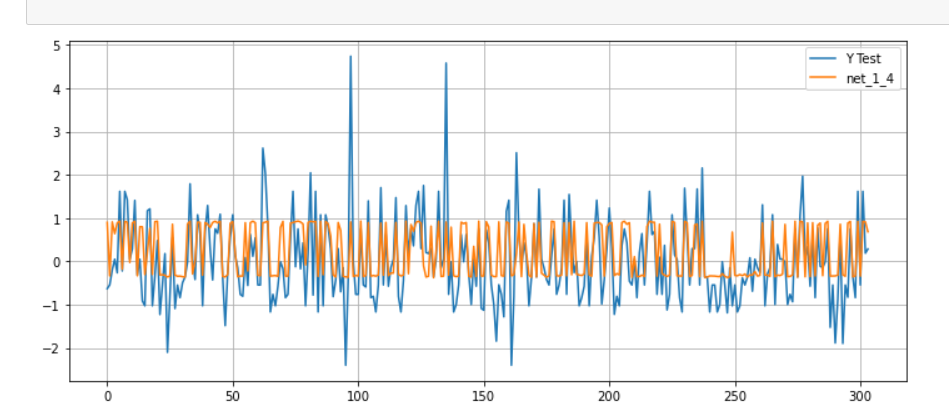


Рис. 33

1. **Итоги исследования**

Рассматриваемые данные имеют достаточно плохую прогнозируемость. Применение моделей прогнозирования на исходных данных не принесло значимого результата. Можно сделать уверенное предположение о том, что исходные данные были аугментированы.

Вероятно, что для решения задач прогнозирования показателей требовался другой алгоритм аугментации, основанный первых 23 строчках основного файла. Однако, в связи с тем, что исходя из условий задачи отсутствуют сведения о возможности такого подхода, в ходе исследования была предпринята попытка создания алгоритма и метрик изменения данных, а также зависимости прогнозируемых показателей от этих метрик.

Также была разработана Flask модель для прогнозирования модуля упругости при растяжении.