Компания «Газпром «ЦПС»

Программа проектов цифрового проектирования пуско-наладочных работ и капитального строительства

Проект «Цифровизация сварочного производства и исполнительной документации» компании»

ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗИ МЕЖДУ ПОКАЗАТЕЛЕМ БРАКА НА ПРОИЗВОДСТВЕ И РАЗЛИЧНЫМИ ПАРАМЕТРАМИ ИСПОЛЬЗОВАННОЙ ДЛЯ ПРОИЗВОДСТВА ТЕХНОЛОГИИ

The Study Of The Relationship Between The Indicator Of Manufacturing Defects And Various Parameters Of The Technology Used For Production: An Empirical Approach.

Эмпирическое исследование для компании «ГАЗПРОМ ЦПС»

СМИРНОВА Владимира Евгеньевича

Руководитель проекта

ТКАЧЕНКО Андрей Николаевич

Санкт-Петербург

2024

Содержание

ВВЕДЕНИЕ

Данная исследовательская работа направлена на проработку **исследовательского вопроса**, который формулируется следующим образом: «Какие технологические факторы в рамках сварочных работ оказывают наибольшее влияние на показатель брака на производстве?»

Исследовательская гипотеза же, в свою очередь, представлена в следующем виде: «Показатель брака на производстве в рамках сварочных работ связан с различными параметрами использованной для производства технологии».

Цель исследования заключается в выявлении закономерностей между факторами сварочных работы и их результатом (уровень брака) для определения способов повышения их качества. Цель исследования может быть достигнута при условии выполнения **следующих задач**:

1. Определить какие факторы (переменные) могут быть получены из имеющегося набора данных для построения регрессионной модели;
2. Сформировать гипотезы для дальнейшей валидации через выделенные факторы;
3. Подготовить данные для построения регрессионной модели;
4. Выбрать спецификацию(и) регрессионной модели(ей);
5. Построить регрессионную(ые) модель(и);
6. Провалидировать выдвинутые гипотезы и проинтерпретировать результаты;
7. Подготовить визуализацию результатов анализа;
8. Сформировать промежуточные и основные выводы исследования.

Объектом исследования являются различные характеристики, параметры и факторы задействованного в рамках сварочных работ стыка.

Предметом исследования является взаимосвязь характеристик использованной для производства технологии и показателем брака на производстве.

Исследование проводилось на выборке, собранной за 9 месяцев сварочных работ на объекте УКПГ-1 Ковыктинского ГКМ, проводимой АО «СтройТрансНефтеГаз».

Методология исследования: эконометрическое моделирование с применением методологии регрессионного анализа с бинарной зависимой переменной (логит-модель регрессии), логистическая регрессия как метод машинного обучения.

Используемое ПО: Excel, надстройка Excel Power Query, Python (Jupiter NoteBook)

Работа состоит из трёх глав, списка использованной литературы и приложения. В первой главе рассматривается сущность регрессионного анализа как мощного аналитического инструмента в контексте трубосварочных работ. Во второй главе дано описание процесса формирование выборки данных, описание переменных, обоснованы и описаны принятые для исследования гипотезы. В третьей главе дано описание методологии исследования, проведенного регрессионного анализа; описан алгоритм анализа выборки с использованием методов машинного обучения, а именно логистической регрессии как одного из основных методов классификации в машинном обучении; дано описание метода эконометрического анализа в контексте бинарной зависимой переменной; представлены основные выводы исследования и предлагаемые организационные решения.

ГЛАВА 1. РЕГРЕССИОННЫЙ АНАЛИЗ КАК МЕТОД ОЦЕНКИ ВЗАИМОСВЯЗЕЙ В ДАННЫХ

1.1. Как регрессионная модель может быть использована на практике при решении производственных задач

Регрессионная модель является основным инструментом для анализа зависимостей между переменными, - переменными, в нашем случае, могут являться факторы производства, особенности используемой в производстве технологии и прочие технологические аспекты. Так, в контексте трубосварочных работ, регрессионные модели позволяют выявлять и количественно оценивать влияние таких технологических параметров, как температура, давление, скорость сварки и другие (данные параметры являются оказывающими непосредственное влияние на зависимую переменную факторами, то есть, независимыми переменными или так называемыми регрессорами), на, например, условное конечное качество сварного соединения (качество, в свою очередь, является зависимой переменной в рамках модели).

Говоря же о практическом применении регрессионного анализа в трубосварочных работах, данный мощный аналитический инструмент предоставляет возможность оптимизации технологического процесса как такового. С его помощью можно не только выявить ключевые факторы, влияющие на качество сварки, оценить силу их влияния в статистическом и количественном выражениях, но и прогнозировать результаты при изменении условий производства.

Далее будут рассмотрены различные спецификации регрессионных моделей с целью получения краткого экскурса в “математику”, лежащую в их основе.

1.1.1. Линейная регрессия

Линейная регрессия - это один из самых простых и популярных методов анализа данных, который помогает выявить и описать взаимосвязь между двумя или более переменными. В основе линейной регрессии лежит предположение, что эта взаимосвязь можно выразить в виде прямой линии.

Представим себе простую ситуацию: мы хотим понять, как такой показатель, как число учеников на одного учителя в классе, влияет на результаты написания тестирования учеником.

У нас есть данные о коэффициенте числа учеников на одного учителя в классе и о результатах написания тестирования учеником. Линейная регрессия поможет нам определить, насколько сильно и как именно независимая переменная влияет на зависимую.

Для того, чтобы дать оценку влияния независимой переменной на зависимую, линейная регрессия строит так называемую "линию наилучшего соответствия" (“линия регрессии”, “линия тренда”) — прямую линию, которая проходит через точки на графике (соответствующие реальным наблюдениям по переменным) и максимально точно отражает общую тенденцию в данных. Эта линия описывается уравнением вида:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

Figure 1 Модель линейной регрессии в базовом виде

Где:

- Индекс i пробегает по всем наблюдениям, in = 1,..., ;
- Yi - зависимая переменная, регрессируемая переменная или просто переменная слева; то есть это то, что мы пытаемся предсказать или объяснить (в нашем примере — результат написания тестирования учеником);
- X1i - независимая переменная, объясняющая переменная, регрессор или просто переменная справа; то есть фактор, который, как мы предполагаем, влияет на Y (в нашем примере — число учеников на одного учителя в классе). Регрессоров в рамках модели может быть более, чем один - зависит от того, какие гипотезы мы хотим проверить, оценивая ту или иную спецификацию модели;
- Bo - это константа линии теоретической регрессии или свободный член. Это значение Y при X = 0; тем не менее, он не всегда имеет какую-либо осмысленную интерпретацию, и оценивания его показатель, всегда стоит ориентироваться на здравый смысл и логику;
- B1 - угловой коэффициент линии теоретической регрессии; то есть, это коэффициент, показывающий, насколько сильно число учеников влияет на результаты написания тестов. Если B1 положительное, это значит, что с ростом отношения учеников на учителя качество написания тестирования

увеличивается, а если отрицательное - уменьшается. Прочие коэффициенты даны для наглядности, ведь переменных, оказывающих влияние на зависимую переменную, может быть много;

- β0 + β1Xi – линия теоретической регрессии, линия регрессии генеральной совокупности, или функция регрессии генеральной совокупности;
- u является компонентой ошибок, которая, как обычно, представляет все другие неучтенные факторы, определяющие зависимую переменную.

После того как мы построили линию наилучшего соответствия, мы можем использовать её для предсказания результатов тестирования на основании отношения числа учеников на одного учителя. Например, если у нас есть новое значение для отношения числа учеников на одного учителя, мы можем подставить его в уравнение и получить предсказание для результатов тестирования [пример из Stock, James H, and Mark W. Watson. Introduction to Econometrics].

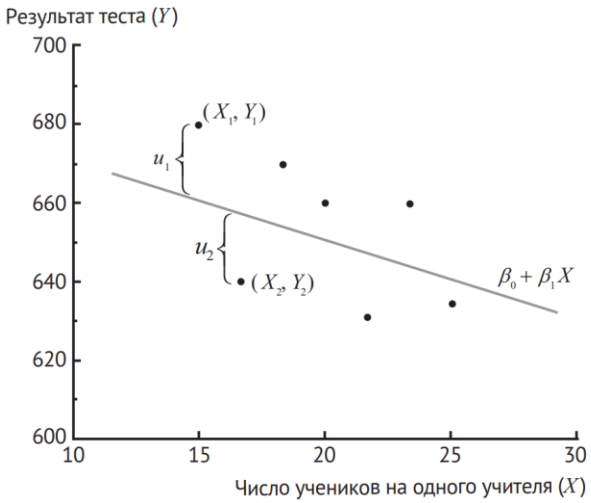

Figure 2 Диаграмма рассеяния результатов тестов

В целом, линейная регрессия - это мощный инструмент для анализа данных, который, несмотря на свою простоту, может быть очень полезен в самых разных задачах, включая анализ процессов в трубосварочных работах.

В нашем исследовании, если бы мы знали теоретические коэффициенты в модели, мы могли бы предсказать средний результат брака или “годности” элемента на основе информации о среднем значениях переменных в рамках различных спецификаций.

Тем не менее, важно помнить, что нам необходимо подобрать такую спецификацию регрессионной модели, которая бы описывала зависимости в наборе данных наилучшим образом, ведь условная линейная регрессия предполагает линейную зависимость между переменными, что не всегда соответствует реальности. Если зависимость между переменными более сложная (например, криволинейная), линейная регрессия может дать неточные результаты; или, если же зависимой переменной является не количественная переменная, выраженная в численной форме, а бинарная переменная (то есть, в формате ДА / НЕТ или же 1 / 0 или же БРАК / НЕ БРАК), - именно здесь на помощь нам придут модели бинарного выбора.

1.1.2. Модели бинарного выбора

Модели бинарного выбора - это особый класс регрессионных моделей, которые используются для анализа и предсказания ситуаций, где результат может принимать одно из двух возможных значений. Такие модели применяются, когда нас интересует вероятность того, что событие произойдет или не произойдет, и часто используются в задачах классификации. Модели бинарного выбора представлены следующими спецификациями:

1. Линейная вероятностная модель
2. Модель Пробит
3. Модель Логит (логистическая регрессия)

1.1.3. Линейная вероятностная модель

Например, линейная вероятностная модель, по сути, представляет собой модель множественной регрессии в случае, когда зависимая переменная является бинарной, а не непрерывной. Поскольку зависимая переменная Y является бинарной, теоретическая функция регрессии соответствует вероятности того, что зависимая переменная равна 1 при заданном значении X. Оценка коэффициента β1 при регрессоре X представляет собой изменение вероятности того, что Y = 1 при единичном изменении[1] X [Stock, James H, and Mark W. Watson. Introduction to Econometrics].

[1] Stock, James H, and Mark W. Watson. Introduction to Econometrics [Электронный ресурс] URL: https://www.sea-stat.com/wp-content/uploads/2020/08/James-H.-Stock-Mark-W.-Watson-Introduction-to-Econometrics-Global-Edition-Pearson-Education-Limited-2020.pdf

Фактически это равносильно тому, что коэффициенты при регрессорах отображают вероятности в рамках изменения зависимой переменной.

Условно, при коэффициенте 0,011 у переменной диаметр можно было бы говорить о том, что увеличение параметра диаметр на 1 мм приводит к увеличению вероятности возникновения **не**бракованной детали на производстве на 0,011%, если наша зависимая переменная “брак” дана в виде 1 - если сварочная операция дала на выходе небракованную деталь, и 0 - если бракованную.

Линейная вероятностная модель является самой простой для построения и интерпретации моделью бинарного выбора, обладающей в то же время и наибольшими изъянами.

Важный недостаток модели следующий - она может предсказывать вероятности, выходящие за пределы допустимого диапазона от 0 до 1. Это происходит потому, что линейная функция не ограничена и может принимать любые значения. Этот недостаток делает данную модель менее популярной по сравнению с другими моделями бинарного выбора, такими как пробит и логит.

1.1.4. Пробит- и логит-модели регрессии

Пробит- и логит-модели представляют собой нелинейные регрессионные модели, специально предназначенные для бинарных зависимых переменных. Поскольку регрессии с бинарной зависимой переменной Y моделируют вероятность того, что Y = 1, то имеет смысл использовать нелинейную постановку модели, которая приводит к тому, что прогнозируемые значения лежат в диапазоне от 0 до 1. Поскольку интегральные функции распределения вероятностей (c. d. f) дают значения вероятностей, расположенные между 0 и 1, то и они используются в логит- и пробит-моделях регрессии. В пробит-модели используется стандартная нормальная функции распределения вероятностей. В логит-модели, которая также называется логистической регрессией, используется логистическая функция распределения [Stock, James H, and Mark W. Watson. Introduction to Econometrics].

Обратим внимание на логит-спецификацию, которая и была использована в нашем исследовании.

Логит-модель или логистическая регрессия

Модель логит, или логистическая регрессия, является одной из самых широко используемых моделей бинарного выбора. В основе модели логит лежит

логистическая функция, которая также ограничивает предсказанные вероятности в диапазоне от 0 до 1. То есть, логистическая регрессия использует логистическую функцию (также называемую сигмоидной функцией) для преобразования линейной комбинации независимых переменных в вероятность.

Логистическая функция описывается следующим уравнением:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Figure 3 Уравнение для логистической функции (сигмоиды)

Где:

- P(Y = 1) — это вероятность того, что зависимая переменная Y примет значение 1;
- B0— свободный член;
- B1, B2, …, Bn — коэффициенты, которые указывают на влияние соответствующих независимых переменных X1, X2, ..., Xn на вероятность исхода;
- X1, X2,..., Xn — независимые переменные.
- e — основание натурального логарифма (приблизительно 2.718).

На графике ниже можно ознакомиться с визуальным представлением логистической функции[2].

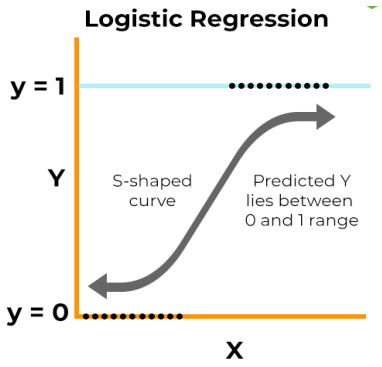

Figure 4 Логистическая регрессия: график в упрощенном виде.

[2] What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices [Электронный ресурс] URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Практический пример работы логистической регрессии

Предположим, мы хотим предсказать вероятность того, что у сварного шва будет дефект, основываясь на двух факторах: температуре сварки (X1) и скорости сварки (X2). Допустим, логистическая регрессия дала следующие коэффициенты:

$$P(Y=1) = \frac{1}{1 + e^{-(2+0{,}03X_1 - 0{,}5X_2)}}$$

- Свободный член (B0).
- Коэффициент перед температурой (B1 = 0.03).
- Коэффициент перед скоростью (B3 = - 0.5).

Теперь, если температура сварки составляет 150°C, а скорость сварки — 3 м/мин, то вероятность дефекта вычисляется так:

$$P(Y=1) = \frac{1}{1 + e^{-(2+0{,}03*150-0{,}5*3)}} \approx \frac{1}{1 + e^{-4{,}35}} \approx 0{,}987$$

То есть вероятность того, что при этих условиях возникнет дефект, составляет около 98,7%.

При этом зависимой переменной может быть как результат как дефект, так и результат как элемент в категории “годен” (что и будет использовано в настоящем исследовании).

Таким образом, логистическая регрессия - мощный инструмент для решения задач бинарной классификации. Она позволяет моделировать и предсказывать вероятность наступления определенного события, учитывая влияние различных факторов.

Выбор спецификации

Выбор в пользу тех или иных спецификаций регрессионных моделей стоит делать основываясь на графическом анализе исходных данных, графическом анализе остатков или же руководствуясь формальными статистическими критериями[3] (условно, Ramsey

[3] Учебники Экономического факультета МГУ. Дружелюбная эконометрика [Электронный ресурс] URL: https://books.econ.msu.ru/Introduction-to-Econometrics/

test для проверки корректности выбранной спецификации исходной модели) [Учебники Экономического факультета МГУ. Дружелюбная эконометрика].

1.2. Методы машинного обучения

Машинное обучение (ML) — это область искусственного интеллекта, которая занимается разработкой алгоритмов и моделей, способных обучаться на данных и делать предсказания или принимать решения без явного программирования. Суть машинного обучения заключается в том, чтобы на основе существующих данных строить модели, которые могут прогнозировать будущие события или находить скрытые закономерности.

Существует множество методов машинного обучения; в данной работе была использована логистическая регрессия, описание которой было дано в предыдущей подглаве.

Логистическая регрессия относится к машинному обучению, поскольку она использует данные для обучения модели, которая может затем делать предсказания на новых, ранее не представленных в выборке данных. Основная цель машинного обучения - это построение моделей, которые могут автоматически выявлять закономерности и принимать решения без необходимости явного программирования. Логистическая регрессия как раз делает это: на основе исторических данных она "учится" выявлять связь между независимыми переменными и вероятностью определенного исхода.

Логистическая регрессия относится к классу **обучения с учителем (supervised learning)**. В задачах с обучением с учителем модель обучается на размеченных данных, где для каждого наблюдения известен правильный результат. В процессе обучения логистическая регрессия строит зависимость между входными переменными (независимыми переменными) и выходной переменной (бинарной целевой переменной). После того как модель обучена, она может предсказывать вероятности для новых данных, определяя, к какому классу они, вероятнее всего, принадлежат.

ГЛАВА 2. ВВОДНЫЕ ЭМПИРИЧЕСКОГО ИССЛЕДОВАНИЯ

2.1. Формирование выборки данных

Данные в рамках выборки относятся к типу панельных данных, полученных в результате наблюдения (наблюдательные данные). Данные описывают характеристики технического характера, связанные с производством.

Excel-файл с данными представляет из себя несколько взаимосвязанных между собой листов. Исходные данные представлены следующими листами:

1. Лист DATA_INITIAL - исходная генеральная совокупность на 61 223 строки, приблизительно 50 столбцов данных, выгрузка данных из 1С в рамках особенностей технологии производства, а также присоединенные к ним столбцы с данными по сотрудникам, которые проводили конкретную сварочную операцию (в исследовании данные по сотрудникам были использованы, однако, в ходе работы анализ по данной выборке был упразднен).
2. Лист DATA_PREPARED - обработанная генеральная совокупность в рамках необходимых предварительных преобразований.

Данные были дополнительно обработаны; в ходе работы на их основе была сформирована выборка, представленная на первом листе Excel-файла:

- Лист TECH_SAMPLE - выборка для оценки технологии производства (в дальнейшем была дополнительно обработана в Python).

По итогам подготовки набора данных был сформирован кодбук для представленной выборки.

2.2. Описание переменных

КОДБУК для выборки “TECH_SAMPLE”

Название переменной	**Тип переменной**	**Описание переменной**	**Принимаемые переменной значения**
date	Категориальная	**Дата**	Используются данные в диапазоне

		Дата проведения сварочного процесса.	2023-2024 гг. в формате ДД.ММ.ГГГГ
name	Качественная	**Фамилия, инициалы бригадира** Фамилия, инициалы бригадира (звеньевого), совершившего процесс.	Принимает значения множественного ФИО в выборке для оценки технологии “TECH_SAMPLE” (ФИО указаны через запятую, если сварочная операция была совершена усилиями нескольких сотрудников).
defect	Количественная, бинарная	**Общий результат** Итоговая характеристика готового элемента в рамках сварочного процесса; переменная, являющаяся зависимой переменной в нашем исследовании	1 - если элемент годен (равносильно значению “годен”); 0 - прочее (равносильно значениям “вырезать”, “превью”, “превью В”, “ремонт”).
temperature	Количественная	**Температура рабочей среды, °C**	Принимает значения от -36 °C

		Температура рабочей среды, при которой была совершена сварочная операция, выраженная в градусах Цельсия	до +29 °C. Описательная статистика: Среднее: 4,6542 Стандартная ошибка: 0,0725 Медиана: 2 Мода: 11 Стандартное отклонение: 15,0513 Минимум: -36 Максимум: 29
diameter	Категориальная	**Наружный диаметр** Наружный диаметр элемента, выраженный в миллиметрах (мм)	Принимает значения от 15 мм до 1220 мм. Описательная статистика (рассчитана как для количественной переменной): Среднее: 154,2732 Стандартная ошибка: 0,7155 Медиана: 108 Мода: 57 Стандартное отклонение: 148,5339

			Минимум: 15 Максимум: 1220
steel_grade	Категориальная	**Марка стали и класс прочности** Марка стали и класс прочности элемента	Принимает следующие качественные значения: • 10 • 20 • 09Г2С • К48 • К52 • К56 • К60
welding_method	Категориальная	**Способ сварки** Используемый при сварочной операции способ сварки	Принимает следующие качественные значения: • ААД • АПГ • РАД • РД • РДАФ
welding_materials	Категориальная	**Сварочные материалы** Используемые при сварочной	Принимает следующие качественные значения (суммарно 14 значений):

		операции сварочные материалы	1. Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Autrod 13.23 классификация ER80S-Ni1по AWS A5.28 2. Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 12.64(2,4 мм) классификация ER70S-6по AWS A5.18 3. Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 13.23(2,4 мм) классификация ER80S-Ni1по AWS A5.28 4. Газовая смесь Ar 80%+CO2 20%; Проволока SUPRAMIG SERIMAX (1,0 мм) классификация

			ER70S-6по AWS A5.18 5. неизвестно 6. ОК 53.70 Ø3.2 7. ОК 53.70 Ø3.2; Проволока ОК Autrod 12.22 (3,0 мм) классификация E7A4-EMпо AWS A5.23; Флюс ОК Flux 10.71 8. ОК 53.70 Ø3.2; Флюс ОК Flux 10.71 9. ОК 53.70 Ø3.2; Электрод ОК 74.70 классификация E8018-G по AWS A5.5 10. Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18 11. Проволока ОК Autrod 12.24 (3,0 мм) классификация F8A2-EA2-A4, F7P0-EA2-A4по AWS A5.23; Флюс ОК Flux 10.71; Электрод ОК

			74.70(3,2 мм) тип Э60 по ГОСТ 9467 12. Электрод ОК 74.70 классификация E8018-G по AWS A5.5 13. Электрод ОК 74.70 тип Э60 по ГОСТ 9467 14. Электрод УОНИ 13/55

2.3. Гипотезы исследования

Гипотеза	**Для какой выборки проверяем**	**Уровень значимости**	**Потенциальный результат тестирования**	**Ожидаемый результат**
H0: Положительная/отрицательная взаимосвязь между предоставленными переменными интереса (Xn: temperature, diameter, steel_grade, welding_metho	TECH_SAMPLE	0,01	Принята / Отвергнута	Существует статистически значимая положительная взаимосвязь

d, welding_materials) и фактом отсутствия / наличия брака (defect) отсутствует. H1: Существует положительная /отрицательная взаимосвязь между предоставленными переменными интереса (Xn: temperature, diameter, steel_grade, welding_method, welding_materials) и фактом отсутствия / наличия брака (defect).				

ГЛАВА 3. ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗИ МЕЖДУ ПОКАЗАТЕЛЕМ БРАКА НА ПРОИЗВОДСТВЕ И РАЗЛИЧНЫМИ ПАРАМЕТРАМИ ИСПОЛЬЗОВАННОЙ ДЛЯ ПРОИЗВОДСТВА ТЕХНОЛОГИИ

3.1. Подготовка выборки данных для статистического анализа

Для начала были загружены используемые в исследовании библиотеки:

```
import msoffcrypto
import pandas as pd
from io import BytesIO
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import statsmodels.api as sm
from statsmodels.stats.diagnostic import het_breuschpagan
```

3.2. Разведочный анализ данных

Для начала были построены столбчатые диаграммы средних значений для количественных и категориальных переменных по среднему числу годных деталей, что позволило в доступной форме сделать базовые оценки представленности различных переменных в рамках выборки, а также их числа по отношению к проценту брака. Напоминание: среднее значение (mean) по колонке defect стремится к 1, если процент брака на производстве снижается, и наоборот, стремится к 0, если повышается. Процент годных деталей определяется по оси Y; процент бракованных деталей - как:

$$\text{Процент бракованных деталей} = \mathbf{1} - \text{Значение по оси } \boldsymbol{Y}$$

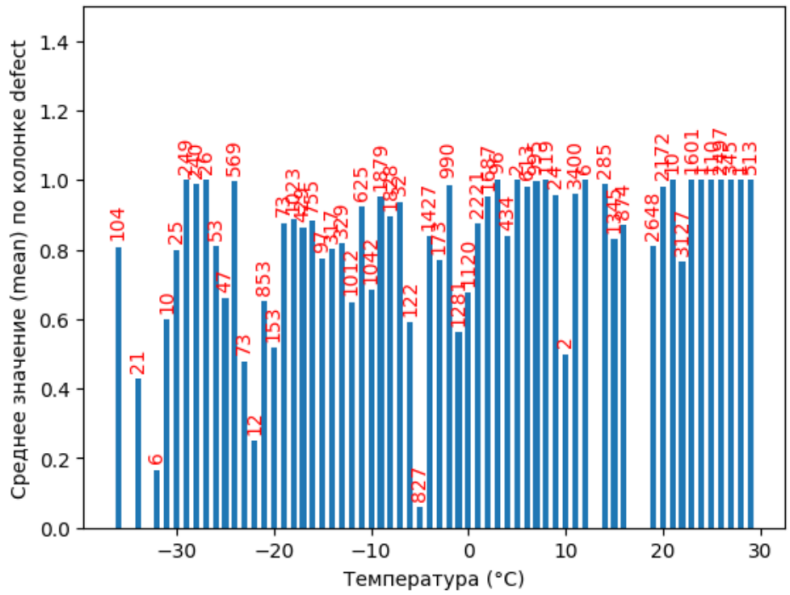

Figure 5 Среднее число (mean) годных деталей (когда переменная defect = 1)

Анализируя столбчатую диаграмму, можно предварительно обратить внимание, что в условном диапазоне температуры -5 градусов находится 827 значений, которым свойственно среднее значение годных деталей на уровне менее, чем 0.1-0.2; то есть, из общей массы в 827 наблюдений менее 10-20% соответствуют трубосварочной операции, результатом которой являлся элемент со статусом “годен”.

Остальные значения температуры либо дают более, чем 50% годных деталей, либо же им присуще слишком малое число наблюдений для релевантных выводов.

Можно предположить, что, возможно, при таких температурах сварщики не прогревают трубы, внутри труб накапливается влага, конденсат, который необходимо прогреть; то есть, требуется обсушить трубу перед сваркой и только после этого сварщик может приступать к процессу сварки. Косвенно, данный аспект может оказывать негативный результат на итоговый результат сварки.

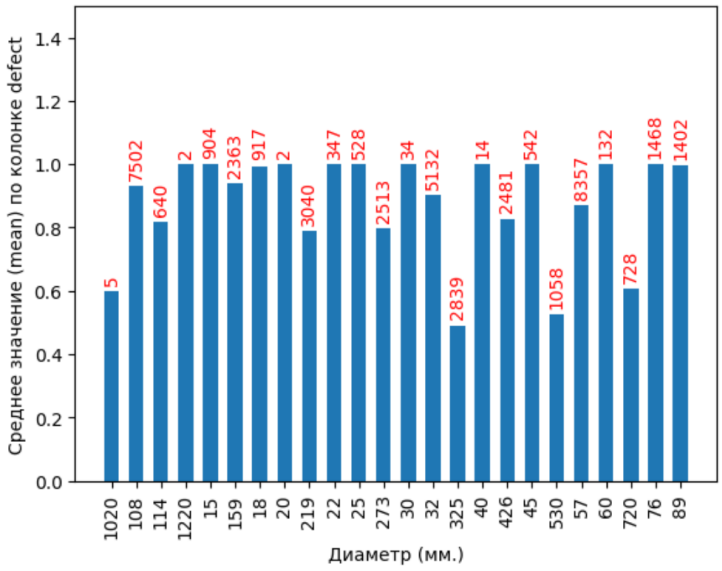

Figure 6 Среднее число (mean) годных деталей (когда переменная defect = 1)

Говоря же о релеватных значениях диаметра, можно заметить, что у значений диаметра трубы 325 мм, 530 мм, 720 мм относительно сравнительно невысокие уровни годности деталей (колеблется в районе 50-60%); остальным в среднем присуще значение на уровне не менее, чем 80%.

Для того, чтобы более подробно оценить данные по переменной, был построен дополнительный, более детальный график в Excel:

Figure 7 Количество уникальных значений для переменной диаметр

А конкретные значения выведены в отдельную таблицу:

Наружный диаметр	Количество элементов с данным наружным диаметром	Количество годных элементов с данным наружным диаметром	Количество бракованных элементов с данным наружным диаметром	Доля годных элементов с данным наружным диаметром
159	2363	2222	141	94,03%
114	640	523	117	81,72%
108	7502	7000	502	93,31%
15	904	904	0	100,00%
219	3040	2404	636	79,08%
89	1402	1397	5	99,64%
57	8357	7270	1087	86,99%
45	542	542	0	100,00%
426	2481	2054	427	82,79%
273	2513	2002	511	79,67%
32	5132	4638	494	90,37%
76	1468	1468	0	100,00%
530	1058	557	501	52,65%
60	132	132	0	100,00%
30	34	34	0	100,00%
325	2839	1394	1445	49,10%
720	728	443	285	60,85%
22	347	347	0	100,00%
25	528	528	0	100,00%
18	917	912	5	99,45%
1220	2	2	0	100,00%
1020	5	3	2	60,00%
40	14	14	0	100,00%
20	2	2	0	100,00%

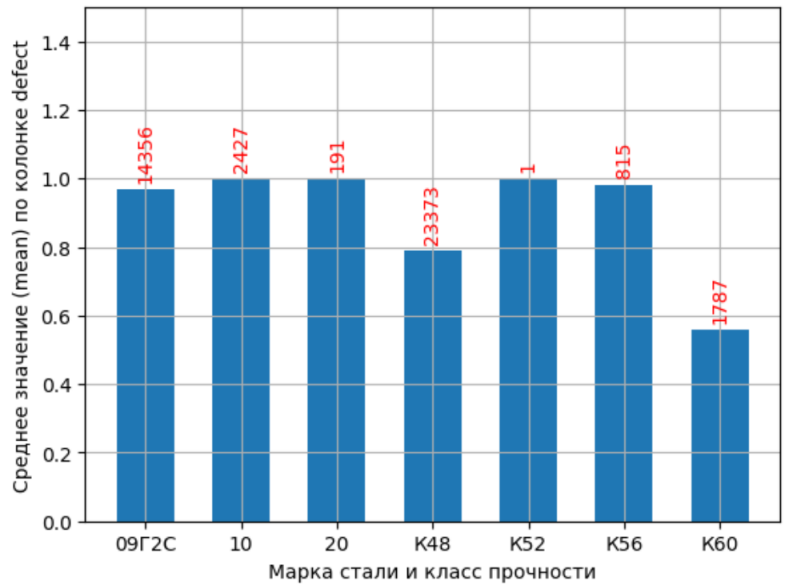

Figure 8 Среднее число (mean) годных деталей (когда переменная defect = 1)

Касаемо марки стали и класса прочности, можно сказать, что марка стали К60 показывает худший результат качества в сравнении с другими, держа показатель годности на уровне около 60%.

Отдельное внимание стоит обратить на марки 09Г2С и К48 - наиболее представленные в выборке марки:

- 09Г2С при количестве наблюдений 14 356 (второе по размеру значение в выборке) показывает годность на уровне ~ 95%;
- К48 же показывает значение ~79% при количестве наблюдений 23 373 (наиболее представленное значение в выборке).

Более рациональные выводы можно было бы делать при приблизительно равном числе наблюдений в рамках каждой марки стали; тем не менее, на данный момент можно сказать, что марка стали К60 показывает себя не очень хорошо с точки зрения итогового качества сварочного стыка; К48 тоже дает не очень благоприятные результаты, так как на самое большое значение наблюдений в рамках марок стали приходится 20% негодных, что соответствует ~4 674,6 некачественным деталям.

С целью того, чтобы более подробно оценить данные по переменной, был построен более детальный график в Excel; отдельно выведена таблица с данными.

Figure 9 Количество уникальных значений по переменной сталь и класс прочности

Марка стали и класс прочности	Количество элементов с данной маркой стали и классом прочности	Количество годных элементов с данной маркой стали и классом прочности	Количество бракованных элементов	Доля годных элементов среди элементов с данной маркой стали и классом прочности
К48	23373	18472	4901	79,03%
09Г2С	14356	13901	455	96,83%
10	2427	2427	0	100,00%
К60	1787	999	788	55,90%
20	191	191	0	100,00%
К56	815	801	14	98,28%
К52	1	1	0	100,00%

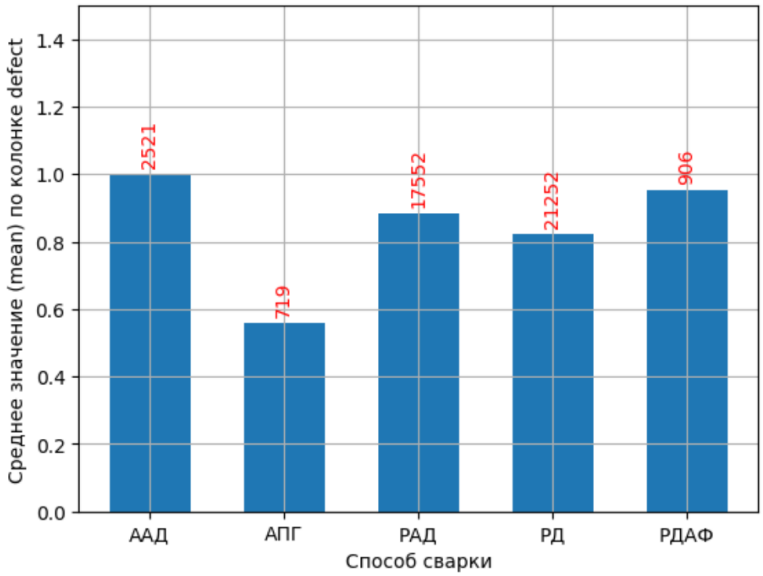

Figure 10 Среднее число (mean) годных деталей (когда переменная defect = 1)

В отношении способа варки можно обнаружить следующие тенденции:

- Способ варки АПНГ показывает себя хуже всего с точки зрения итогового результата сварочной операции, показывая годность на уровне ~ 55-57%; тем не менее, именно с помощью этого способа сварки было совершено меньше всего сварочных операций;
- Способ варки ААД показал наилучший показатель, выдав ~ 100% годных элементов;
- РДАФ - ~ 95% годных элементов.
- Наиболее представленные способы сварки в выборке - РД и РАД, показали уровни ~ 82% и ~ 87% соответственно.

Для того, чтобы более подробно оценить данные по переменной, снова был построен дополнительный график в Excel; выведена таблица с данными.

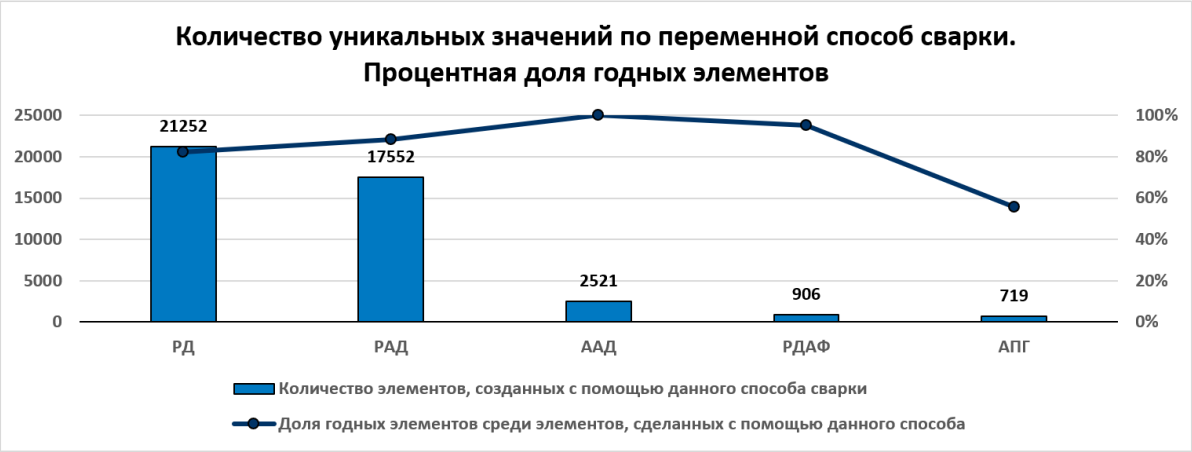

Figure 11 Количество уникальных значений по переменной способ сварки

Использованный способ сварки	Количество элементов, созданных с помощью данного способа сварки	Количество годных элементов, сделанных с помощью данного метода	Количество бракованных элементов	Доля годных элементов среди элементов, сделанных с помощью данного способа
РД	21252	17491	3761	82,30%
РАД	17552	15516	2036	88,40%
ААД	2521	2521	0	100,00%
РДАФ	906	863	43	95,25%
АПГ	719	401	318	55,77%

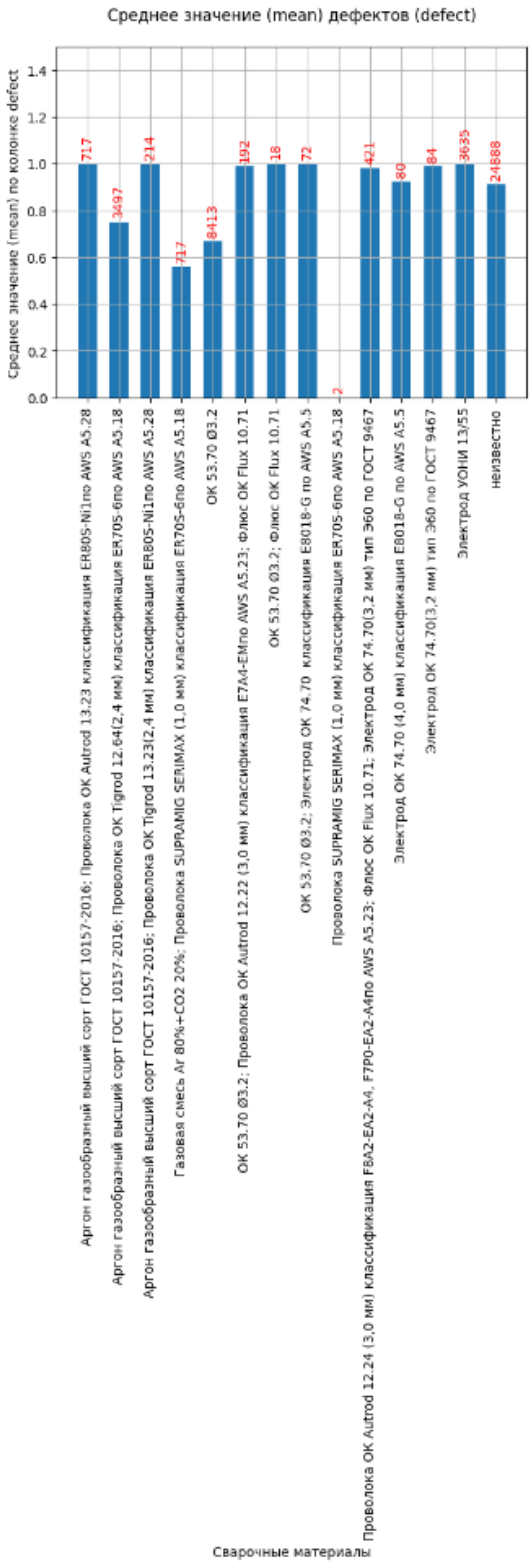

Figure 12 Среднее число (mean) годных деталей (когда переменная defect = 1)

Говоря же о сварочных материалах, то значения годности элемента разнятся; отдельно можно выделить группу “неизвестно”, которой присуще 24 888 наблюдений, что делает ее наиболее представленной группой материалов в выборке. Это говорит о том,

что данные по используемым сварочным материалам не всегда заполняются в системах хранения данных.

Дополнительный график в Excel и таблица с данными представлены ниже.

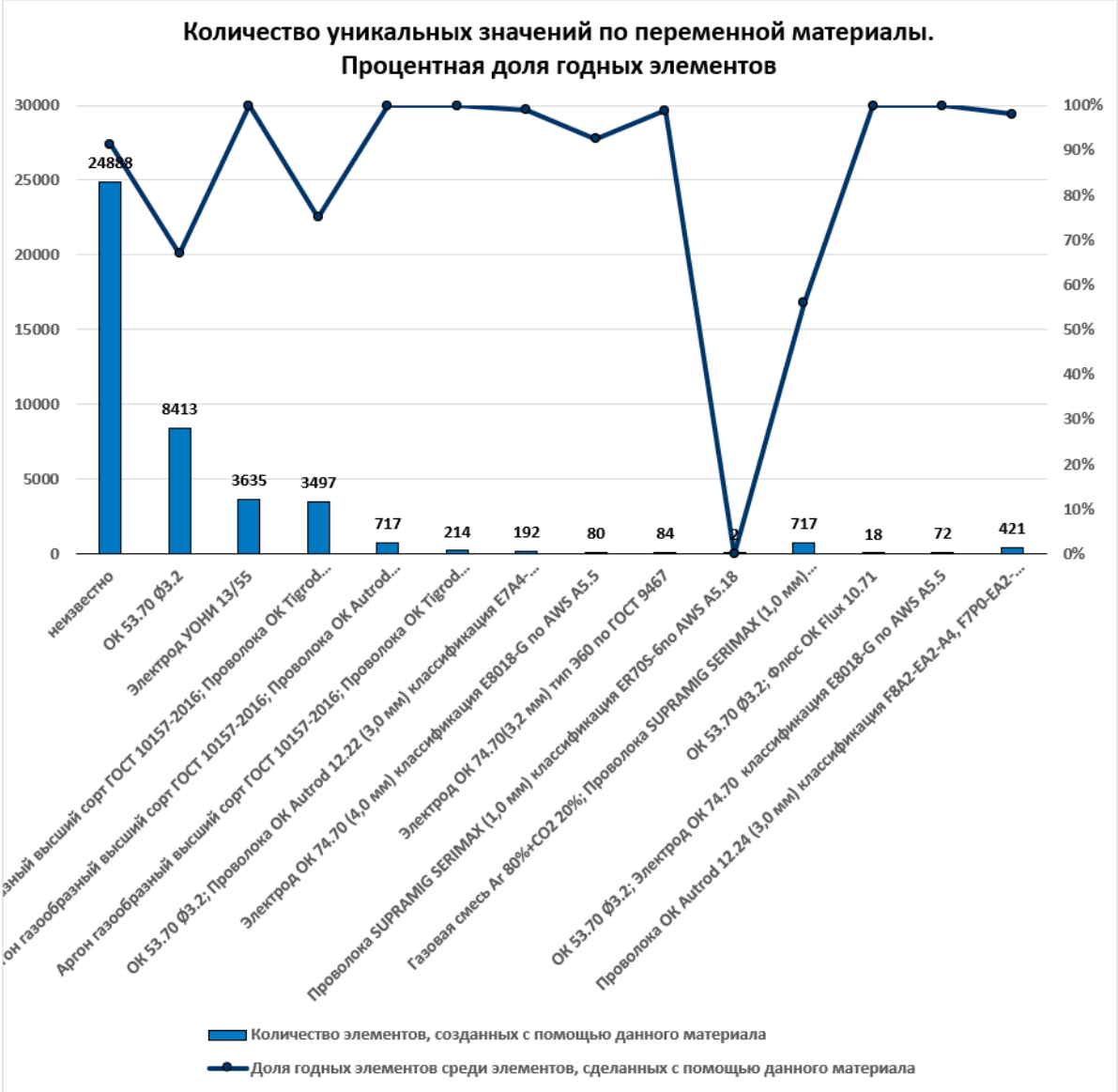

Figure 13 Количесвто уникальных значений по переменной материалы

Использованный материал	Количество элементов, созданных с помощью данного материала	Количество годных элементов, сделанных с помощью данного материала	Количество бракованных элементов	Доля годных элементов среди элементов, сделанных с помощью данного материала
неизвестно	24888	22719	2169	91,28%
ОК 53.70 Ø3.2	8413	5631	2782	66,93%
Электрод УОНИ 13/55	3635	3635	0	100,00%
Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 12.64(2,4 мм) классификация ER70S-6по AWS A5.18	3497	2625	872	75,06%
Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Autrod 13.23 классификация ER80S-Ni1по AWS A5.28	717	717	0	100,00%
Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 13.23(2,4 мм) классификация ER80S-Ni1по AWS A5.28	214	214	0	100,00%
ОК 53.70 Ø3.2; Проволока ОК Autrod 12.22 (3,0 мм) классификация E7A4-EMпо AWS A5.23; Флюс ОК Flux 10.71	192	190	2	98,96%
Электрод ОК 74.70 (4,0 мм) классификация E8018-G по AWS A5.5	80	74	6	92,50%
Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	84	83	1	98,81%
Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18	2	0	2	0,00%
Газовая смесь Ar 80%+CO2 20%; Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18	717	401	316	55,93%
ОК 53.70 Ø3.2; Флюс ОК Flux 10.71	18	18	0	100,00%
ОК 53.70 Ø3.2; Электрод ОК 74.70 классификация E8018-G по AWS A5.5	72	72	0	100,00%
Проволока ОК Autrod 12.24 (3,0 мм) классификация F8A2-EA2-A4, F7P0-EA2-A4по AWS A5.23; Флюс ОК Flux 10.71; Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	421	413	8	98,10%

Помимо этого, для каждого метода сварки были определены используемые в рамках метода материалы.

Метод сварки	Используемые материалы	Количество используемых материалов в рамках метода сварки
ААД	неизвестно	2521
АПГ	Газовая смесь Ar 80%+CO2 20%; Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18	717
	Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS	2

	A5.18	
РАД	Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Autrod 13.23 классификация ER80S-Ni1по AWS A5.28	717
	Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 12.64(2,4 мм) классификация ER70S-6по AWS A5.18	3497
	Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 13.23(2,4 мм) классификация ER80S-Ni1по AWS A5.28	214
	неизвестно	13124
РД	ОК 53.70 Ø3.2	8413
	ОК 53.70 Ø3.2; Электрод ОК 74.70 классификация E8018-G по AWS A5.5	72
	Электрод ОК 74.70 (4,0 мм) классификация E8018-G по AWS A5.5	80

	Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	84
	Электрод УОНИ 13/55	3635
	неизвестно	8968
РДАФ	ОК 53.70 Ø3.2; Проволока ОК Autrod 12.22 (3,0 мм) классификация E7A4-EMпо AWS A5.23; Флюс ОК Flux 10.71	192
	ОК 53.70 Ø3.2; Флюс ОК Flux 10.71	18
	Проволока ОК Autrod 12.24 (3,0 мм) классификация F8A2-EA2-A4, F7P0-EA2-A4по AWS A5.23; Флюс ОК Flux 10.71; Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	421
	неизвестно	275

Описательная статистика

Была построена таблица описательных статистик для количественной переменной “температура” и категориальной “диаметр”. Основные аспекты таблицы представлены в выжимке ниже.

Статистика	Temperature	Diameter
Count (число наблюдений)	42950.00	42950.00

Mean (среднее значение)	4.71	154.70
Std (стандартное отклонение)	15.05	148.59
Min (минимальное значение)	-36.00	15.00
Квантиль уровня 25% (первый квартиль)	-8.00	57.00
Квантиль уровня 50% (медиана)	2.00	108.00
Квантиль уровня 75% (третий квартиль)	19.00	219.00
Max (максимальное значение)	29.00	1220.00

Основные выводы по переменной temperature:

- Колонка temperature варьируется от *-36°C* до *29°C* (большой разброс значений);
- Среднее значение температуры (mean) составляет *4.71°C*, то есть в большинстве случаев ниже *5°C*. В данной выборке есть несколько значений, которые тянут среднее значение вверх, но большинство значений сосредоточено в более низком диапазоне;
- Стандартное отклонение (std) составляет *15.05°C*. Значительный разброс температур в выборке. Температуры могут варьироваться достаточно широко вокруг среднего значения;

- Q1 (1 квартиль) составляет *-8.00°C*, Q3 (3 квартиль) составляет *19°C*. В данном интервале находятся 50% наблюдений, что указывает на довольно широкий разброс;
- Минимальное значение (*-36°C*) и максимальное (*29°C*) показывают наличие экстремальных температур. Данные значения могут быть выбросами;
- В 10% наблюдений температура не превышает *-16°C*, что может указывать на холодные условия;
- Распределение может быть скошенным вправо из-за более высоких значений, которые тянут среднее значение вверх.

Основные выводы по переменной diameter:

- Колонка diameter варьируется от *15 мм.* до *1220 мм.* (большой разброс значений);
- Среднее значение диаметра (mean) составляет *154.7 мм.*, которое является больше медианы *108 мм.*. Как и в случае с температурой: большинство значений в выборке являются ниже среднего, но есть также есть значения, которые являются большими;
- Стандартное отклонение (std) составляет *148.59 мм.*. Значительный разброс диаметров в выборке. Максимальное значение диаметра (1220) значительно превышает среднее значение (mean);
- Q1 (1 квартиль) составляет *57 мм.*, Q3 (3 квартиль) составляет *219 мм.*. В данном интервале находятся 50% наблюдений, что указывает на разнообразие в размерах;
- Минимальное значение (*15 мм.*) и максимальное (*1220 мм.*) говорят о наличии значительных выбросов, особенно высокое значение (*1220 мм.*), которое превышает 75-й процентиль (*219 мм.*);
- Высокие значения в 90-ом и 95-ом процентилях показывают, что есть много объектов с большими диаметрами труб;
- Распределение может быть скошенным вправо, так как среднее значение (*154.7 мм.*) выше медианы (*108 мм.*);
- Часто встречаемый диаметр трубы - *57 мм.*

Оценка нормальности распределения

Были проведены тесты Шапиро-Уилка[4], Колмогорова-Смирнова[5] и Андерсона-Дарлинга[6] на оценку нормальности распределения количественных данных в выборке.

Тесты данного рода являются полезными для проверки распределения случайной величины на нормальность (или же о принадлежности выборки некоторому закону распределения), которая является распространенным предположением, используемым во многих статистических тестах, включая регрессию, дисперсионный анализ, t-тесты и многие другие[7].

Тем не менее, сам факт нормальности распределения регрессоров не является одним из предположений для проведения логистической регрессии; в нашем случае, мы просто проверяем данные на соответствие или несоответствие нормальности в рамках предварительного анализа.

В случае теста Шапиро-Уилка и Колмогорова-Смирнова, если p-значение меньше уровня значимости (чаще всего уровень значимости принимается как 0.05), нулевая гипотеза о нормальности распределения отклоняется, и считается, что данные не распределены нормально.

В тесте Андерсона-Дарлинга результаты включают критические значения для различных уровней значимости; если статистика теста превышает критическое значение, данные не соответствуют нормальному распределению.

Вначале были выведены графики для визуальной оценки нормальности распределения для переменной “температура”.

---

[4] Критерий Шапиро-Уилка [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A8%D0%B0%D0%BF%D0%B8%D1%80%D0%BE-%D0%A3%D0%B8%D0%BB%D0%BA%D0%B0

[5] Критерий Колмогорова-Смирнова [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%9A%D0%BE%D0%BB%D0%BC%D0%BE%D0%B3%D0%BE%D1%80%D0%BE%D0%B2%D0%B0-%D0%A1%D0%BC%D0%B8%D1%80%D0%BD%D0%BE%D0%B2%D0%B0

[6] Критерий Андерсона-Дарлинга [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%90%D0%BD%D0%B4%D0%B5%D1%80%D1%81%D0%BE%D0%BD%D0%B0-%D0%94%D0%B0%D1%80%D0%BB%D0%B8%D0%BD%D0%B3%D0%B0

[7] Как выполнить тест Андерсона-Дарлинга в Python [Электронный ресурс] URL: https://www.codecamp.ru/blog/anderson-darling-test-python/

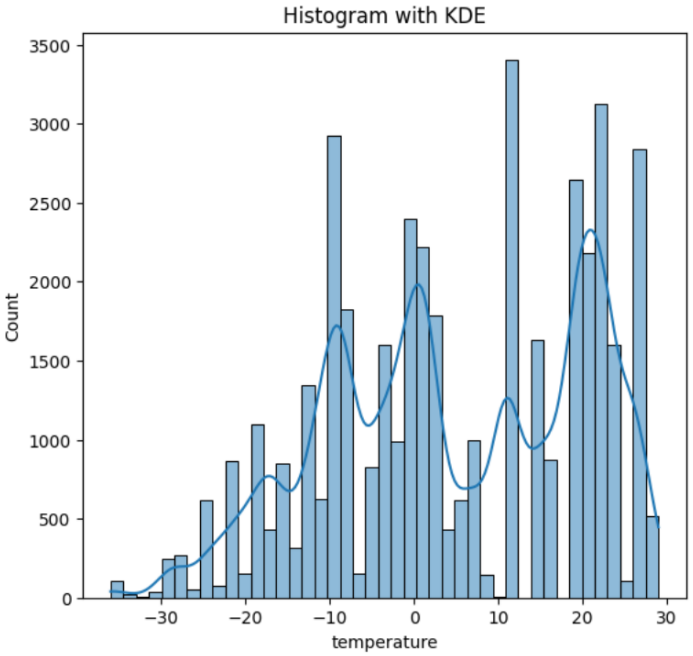

Figure 14 Гистограмма данных по температуре с наложением оценки плотности ядра (KDE, Kernel Density Estimate)

- Судя по графику, данные распределены неравномерно; видны несколько пиков (бимодальное или мультимодальное распределение), что может свидетельствовать о наличии различных подгрупп в данных;
- Наблюдается что-то похожее визуально на левостороннюю асимметрию распределения[8];
- На гистограмму были наложены ядерные оценки плотности; ядерные оценки плотности - способ получения графика плотности в виде непрерывной кривой, то есть, KDE линия сглаживает данные, показывая вероятное распределение значений. В нашем случае она демонстрирует, что есть несколько явно выраженных зон с высокой плотностью значений.

[8] Асимметрия и эксцесс эмпирического распределения [Электронный ресурс] URL: http://mathprofi.ru/asimmetriya_i_excess.html

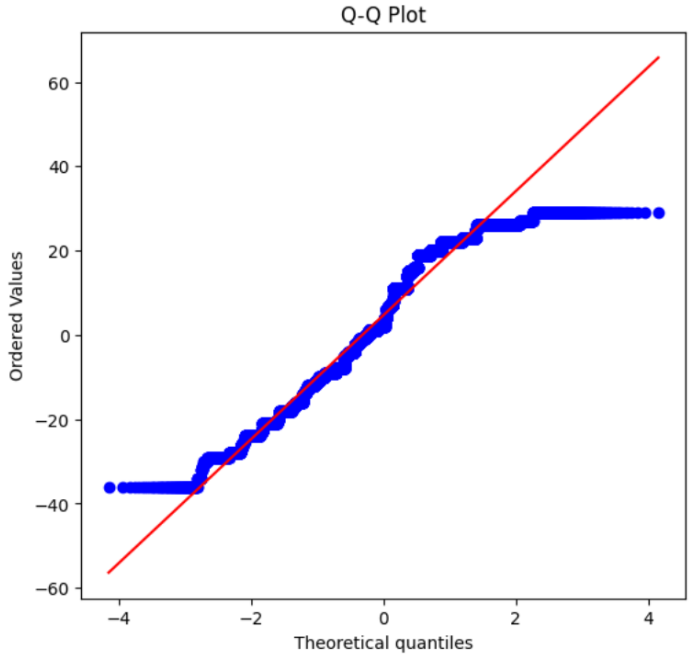

Figure 15 График "Квантиль-квантиль" для переменной "температура"

На графике квантиль-квантиль (или Q – Q график; англ.: Q-Q Plot), также называемом графиком квантилей, можно зрительно оценить подгонку теоретического распределения наблюдаемым данным, т.е. это позволяет нам понять с одного взгляда, какое может быть максимальное отклонение без угрозы невозврата к историческому значению. Q - Q график (Q - квантиль) - это график, на котором квантили из двух распределений расположены относительно друг друга. Чем ближе точки на графике к диагональной прямой, тем ближе распределение исследуемой переменной к нормальному закону. Квантиль-квантиль график, представляет собой инструмент, который помогает нам оценить правдоподобность отклонения среднего от теоретического распределения. На квантиль-квантиль графике показана связь между наблюдаемыми значениями переменных и теоретическими квантилями. Если наблюдаемые значения попадают на прямую линию, то теоретическое распределение хорошо подходит к наблюдаемым данным[9].

В нашем случае же наблюдаются следующие аспекты:

- В нижней части графика точки отклоняются выше линии, а в верхней — ниже, что говорит о том, что хвосты распределения данных длиннее, чем в нормальном распределении (распределение данных имеет тяжелые хвосты);

[9] ГРАФИКИ КВАНТИЛЬ-КВАНТИЛЬ (Q – Q) ДЛЯ ГИДРОЛОГО-ГИДРОЛОГИЧЕСКИХ ДАННЫХ МОЛДОВЫ [Электронный ресурс] URL: https://ibn.idsi.md/sites/default/files/imag_file/p-75-78_3.pdf

- В середине наблюдается довольно хорошее соответствие, но значительные отклонения в крайних значениях говорят о том, что данные все же не являются нормальными;
- Линия данных не является прямой, а скорее имеет изогнутую форму, что также указывает на отклонение от нормальности. Это свидетельствует о том, что данные могут быть скошенными или мультимодальными, как было видно на предыдущей гистограмме;
- Распределение данных не является нормальным. Возможна значительная асимметрия или присутствие нескольких мод (пиков). Это подтверждается предыдущей гистограммой с несколькими пиками.

Результаты тестирования на нормальность же показали следующие результаты:

Тест Шапиро-Уилка:

```
ShapiroResult(statistic=np.float64(0.9525772111149112),
pvalue=np.float64(4.3971614080854674e-76))
```

- statistic: 0.9525 — это значение статистики теста. Значение близкое к 1 указывает на нормальное распределение;
- p-value: 4.39e-76 — это значение p-value. Оно чрезвычайно малое, что указывает на отклонение от нормальности (поскольку p-value намного меньше типичного уровня значимости в 0.05);
- Вывод: данные не соответствуют нормальному распределению, так как p-value слишком малое (менее 0.05).

Тест Андерсона-Дарлинга:

```
AndersonResult(statistic=np.float64(677.1246224207061),
        critical_values=array([0.576, 0.656, 0.787, 0.918, 1.092]),
        significance_level=array([15. , 10. ,  5. ,  2.5,  1. ]),
        fit_result=  params: FitParams(loc=np.float64(4.707823050058207),
                                       scale=np.float64(15.04743984188451))
```

- statistic: 677.1246 — это значение статистики теста. Чем больше это значение, тем больше отклонение от нормального распределения;
- critical_values: массив критических значений для разных уровней значимости (например, 15%, 10%, 5%, и т.д.);

- fit_result: параметры, по которым распределение было подогнано: loc (среднее) и scale (стандартное отклонение);
- success: тест был успешно выполнен;
- Вывод: статистика теста (677.1246) значительно превышает критические значения для всех уровней значимости. Это также указывает на то, что данные не соответствуют нормальному распределению.

Тест Колмогорова-Смирнова:

```
KstestResult(statistic=np.float64(0.13215747355593443),
             pvalue=np.float64(0.0),
             statistic_location=np.int64(19),
             statistic_sign=np.int8(-1))
```

- statistic: 0.1321 — это значение статистики теста, указывающее на максимальное отклонение между фактическим и нормальным распределением;
- p-value: 0.0 — значение p-value, которое здесь равно ~ 0, что указывает на значительное отклонение от нормальности;
- statistic_location: 19 — индекс в данных, где обнаружено максимальное отклонение;
- statistic_sign: -1 — знак, указывающий, в каком направлении отклоняются данные (здесь -1);
- Вывод: p-value равен ~ 0, что указывает на явное отклонение от нормального распределения.

Результаты всех трех тестов (Шапиро-Уилк, Андерсона-Дарлинга и Колмогорова-Смирнова) показывают, что наши данные по переменной “температура” не соответствуют нормальному распределению, так как p-value везде является значительно ниже порогового уровня.

Аналогичные операции были проведены и для переменной диаметр:

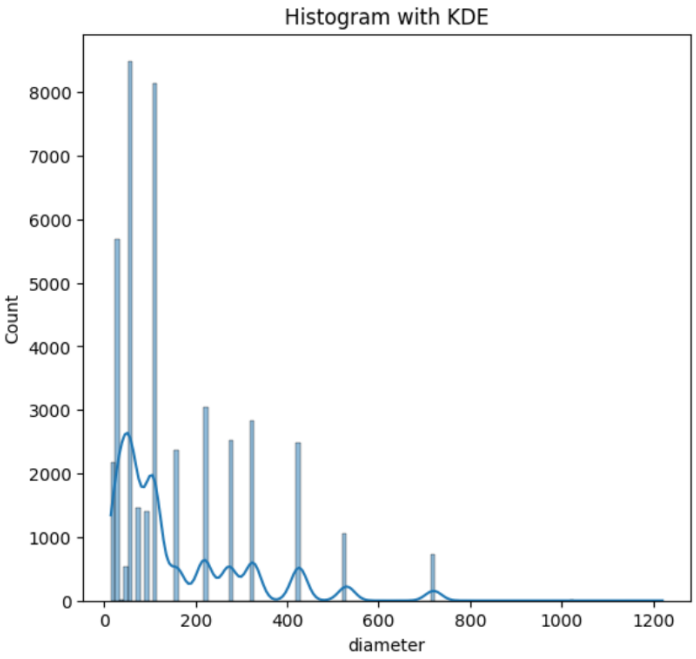

Figure 16 Гистограмма данных по диаметру с наложением оценки плотности ядра (KDE, Kernel Density Estimate)

- Распределение данных асимметрично, со смещением вправо (правосторонняя асимметрия). Это означает, что большинство наблюдений имеют относительно небольшие значения диаметра, а значения, значительно превышающие среднее, встречаются реже;
- Наблюдается несколько локальных максимумов (мод), что указывает на мультимодальное распределение. Это может свидетельствовать о наличии нескольких подгрупп в данных с различными характеристиками диаметра;
- Данные имеют достаточно высокую дисперсию, что отражает большой разброс значений диаметра;
- Наличие нескольких высоких столбиков в правой части графика может указывать на наличие выбросов – значений, значительно отличающихся от основной массы данных.

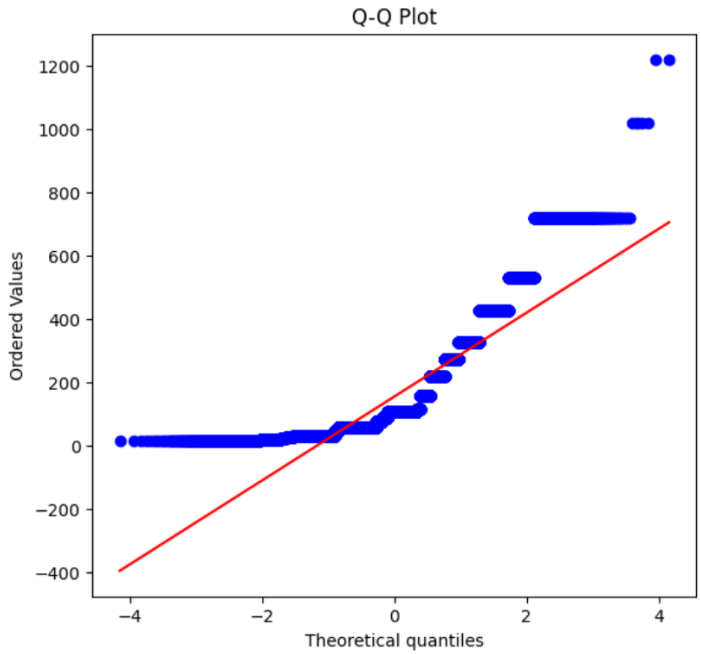

Figure 17 График "Квантиль-квантиль" для переменной "диаметр"

- Аналогично предыдущей диаграмме, можно заметить, что данные не следуют нормальному распределению;
- Хвосты распределения наших данных тяжелее, чем у нормального распределения. Это означает, что у нас есть больше экстремальных значений (как очень больших, так и очень маленьких), чем ожидалось бы при нормальном распределении.
- Зависимость между теоретическими и эмпирическими квантилями нелинейная. Это также указывает на отклонение от нормальности.

Результаты тестирования на нормальность же показали следующие результаты:

Тест Шапиро-Уилка:

```
ShapiroResult(statistic=np.float64(0.793439860318876),
pvalue=np.float64(2.486749574262856e-112))
```

- statistic: 0.7934 — это значение статистики теста. Значение ниже 1 указывает на отклонение от нормальности;
- p-value: 2.49e-112 — p-value чрезвычайно мал (практически равен 0), что говорит о сильном отклонении от нормального распределения;
- Вывод: данные о диаметре не соответствуют нормальному распределению, так как p-value очень маленькое (меньше 0.05).

Тест Андерсона-Дарлинга:

```
AndersonResult(statistic=np.float64(2982.086997283339),
               critical_values=array([0.576, 0.656, 0.787, 0.918, 1.092]),
               significance_level=array([15. , 10. ,  5. ,  2.5,  1. ]),
               fit_result=  params: FitParams(loc=np.float64(154.70249126891736),
                                              scale=np.float64(148.58684760167333))

success: True
message: '`anderson` successfully fit the distribution to the data.'
```

- statistic: 2982.087 — значение статистики теста, которое значительно превышает критические значения;
- critical_values: массив критических значений для уровней значимости (15%, 10%, 5%, 2.5%, и 1%). Например, для уровня значимости 5% критическое значение — 0.787;
- fit_result: параметры распределения, к которым данные были подогнаны: среднее (loc = 154.7) и стандартное отклонение (scale = 148.59);
- Вывод: статистика теста (2982.087) намного выше всех критических значений, что указывает на сильное отклонение от нормального распределения.

Тест Колмогорова-Смирнова:

```
KstestResult(statistic=np.float64(0.2585394366696479),
             pvalue=np.float64(0.0),
             statistic_location=np.int64(108),
             statistic_sign=np.int8(1))
```

- statistic: 0.2585 — максимальное отклонение между распределением диаметра и теоретическим нормальным распределением;
- p-value: 0.0 — p-value равно 0, что указывает на то, что данные сильно отклоняются от нормального распределения;
- statistic_location: 108 — индекс в данных, где было обнаружено максимальное отклонение;
- statistic_sign: 1 — указывает направление отклонения данных (здесь отклонение в положительную сторону);
- Вывод: значение p-value равно ~ 0, что подтверждает отклонение данных о диаметре от нормального распределения.

Таким образом, все три теста показывают, что переменная диаметр не соответствует нормальному распределению.

Анализ на нетипичные данные (выбросы)

Для анализа набора данных на выбросы была построена коробчатая диаграмма[10] (“ящик с усами”; англ.: boxplot) для переменных temperature, diameter:

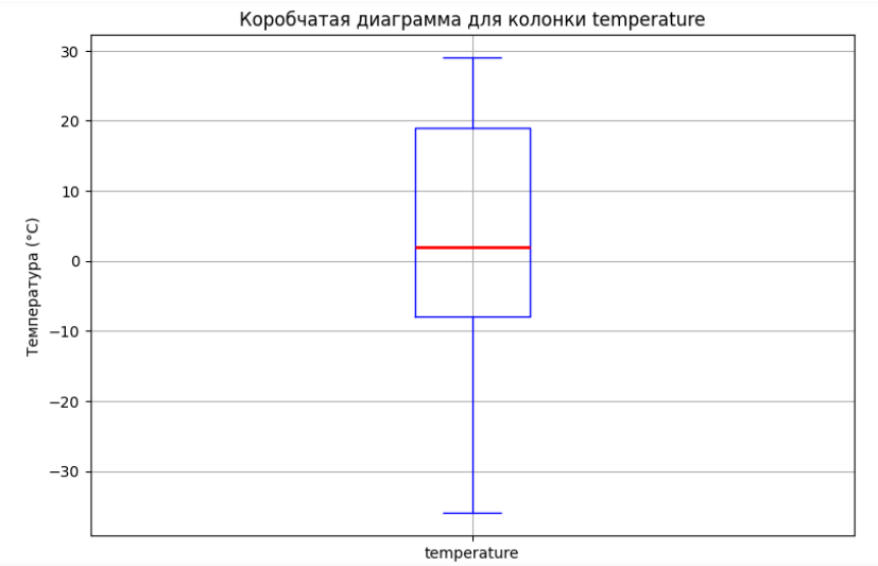

Figure 18 Диаграмма "Ящик с усами" для переменной "temperature"

Интерпретация ящика с усами для переменной “температура”:

- Усы - самые длинные вертикальные линии. Они показывают минимальное и максимальное значения температуры в наборе данных;
- Коробка - вертикальный прямоугольник. Он охватывает значения от первого квартиля (25% данных ниже этого значения) до третьего квартиля (75% данных ниже этого значения). В нашей диаграмме коробка находится примерно между -10°C и 20°C. Это означает, что центральные 50% всех температурных значений находятся в этом диапазоне;
- Красная линия внутри коробки - это медиана. Это значение, которое делит данные на две равные части. Здесь медиана находится несколько выше 0°C; ориентируясь на то, что коробка немного смещена вверх относительно нуля - это снова может указывать на то, что распределение температур слегка положительно скошено, то есть более высоких температур в центральном 50%

[10] Диаграмма размаха ("ящик с усами") [Электронный ресурс] URL: https://datavizcatalogue.com/RU/metody/diagramma_razmaha.html

диапазоне было несколько больше, чем низких; то есть, заметна некоторая асимметрия в распределении температурных значений;

- Отсутствие очень длинных усов говорит о том, что в данных, вероятно, нет сильных выбросов (экстремальных значений), которые значительно отличались бы от остальных; об этом же говорит отсутствие точек единичных данных на диаграмме.

Аналогичный ящик с усами, но уже для переменной диаметр:

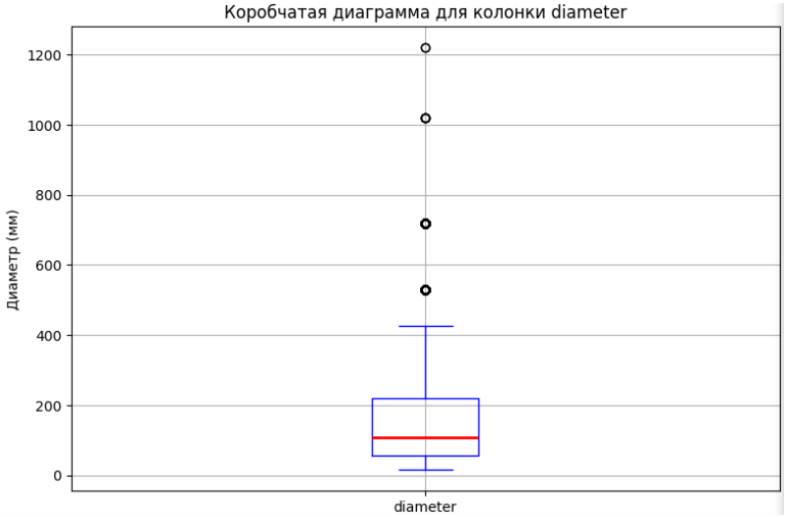

Figure 19 Диаграмма "Ящик с усами" для переменной "diameter"

Интерпретация ящика с усами для переменной “диаметр”:

- Медиана (красная линия внутри коробки) находится примерно на уровне 100 мм. Это означает, что примерно половина объектов имеет диаметр меньше этого уровня, а примерно половина — больше;
- Коробка, которая охватывает значения от первого до третьего квартиля, относительно небольшая. Это говорит о том, что центральные 50% данных сгруппированы вокруг медианы;
- На диаграмме видно четыре точки, расположенные значительно выше коробки. Теоретически, это выбросы - значения, сильно отличающиеся от основной массы данных; тем не менее, в нашем случае, руководствуясь рациональным понимаем наших данных, можно прийти к выводу, что это лишь наименее представленные значения диаметра в нашей выборке; исключать их из анализа полностью, посчитав выбросами, было бы не совсем правильно;
- Диаграмма слегка положительно скошена.

Были также построены графики-гистограммы плотности распределения:

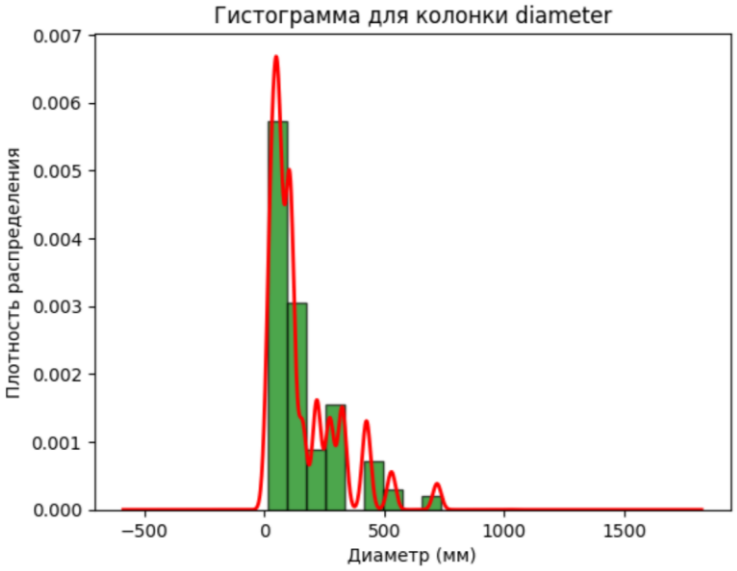

Figure 20 Гистограмма для колонки "diameter". Еще один формат представления

Говоря о таком графике для переменной диаметр, можно сказать, что:

- Распределение данных явно скошено вправо. Это означает, что большинство значений сосредоточено в левой части графика, а длинный "хвост" тянется вправо. Это говорит о том, что есть небольшое количество значений, значительно превышающих среднее;
- Распределение, скорее всего, многомодальное. Видно несколько пиков, что указывает на то, что в данных присутствует несколько групп значений.

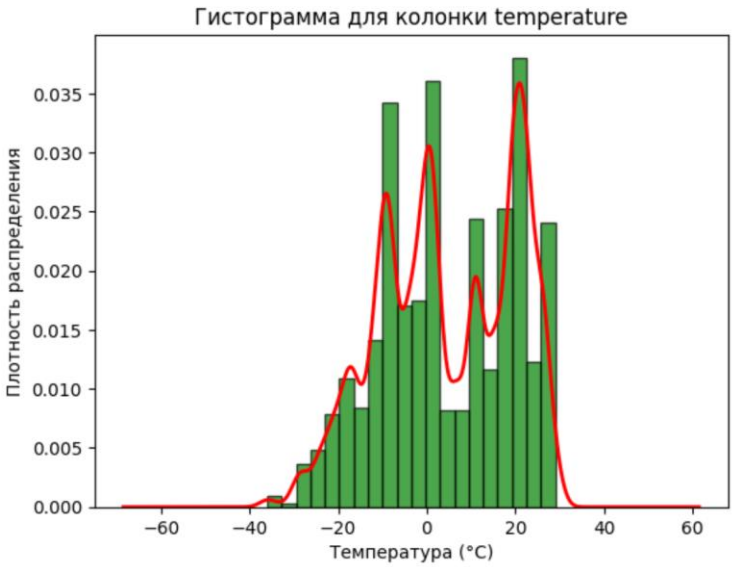

Figure 21 Гистограмма для колонки "temperature". Еще один формат представления

Говоря же о таком графике для переменной температура, можно сказать, что:

- Распределение температуры скошено влево. Это значит, что большинство значений температуры сосредоточено в правой части графика, а длинный "хвост" тянется влево. Это говорит о том, что есть небольшое количество значений, значительно превышающих среднюю температуру;
- Распределение, скорее всего, многомодальное. Видно несколько пиков, что может указывать на наличие нескольких температурных режимов или на то, что данные собирались в разных условиях.

Помимо этого, были построены точечные диаграммы для переменных temperature и diameter. Точечная диаграмма показывает распределение значений температуры по набору данных. Каждая точка на графике соответствует одному наблюдению (строке в данных) и отражает значение температуры для этого наблюдения.

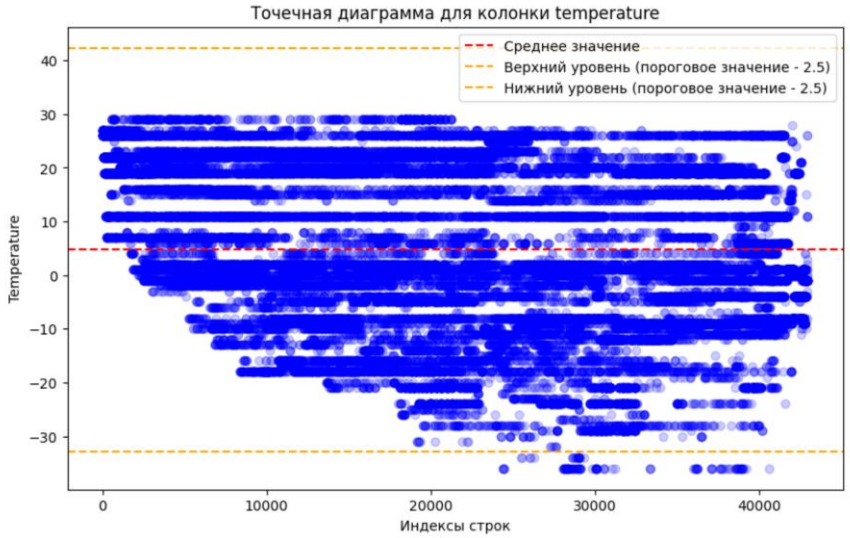

Figure 22 Точечная диаграмма для колонки "temperature"

Точечная диаграмма для переменной температура показала, что:

- В среднем температура в данных ближе к положительным значениям;
- Значения температуры достаточно разбросаны вокруг среднего значения. Это свидетельствует о большой вариабельности температуры в данных;
- На графике нет явных выбросов (значений, сильно отличающихся от остальных);

- На графике обозначены верхний и нижний пороговые значения, равные среднему значению плюс-минус 2.5. Это может использоваться для выделения значений, которые отклоняются от среднего более чем на 2;
- Распределение температуры, скорее всего, скошено влево. Это означает, что большая часть значений сосредоточена в правой части графика (ниже среднего), а хвост распределения тянется влево (к более высоким температурам);
- На графике сложно определить точное количество мод (пиков), но можно предположить, что распределение может быть многомодальным, то есть иметь несколько пиков. Это может указывать на наличие нескольких подгрупп данных с различными температурными режимами.

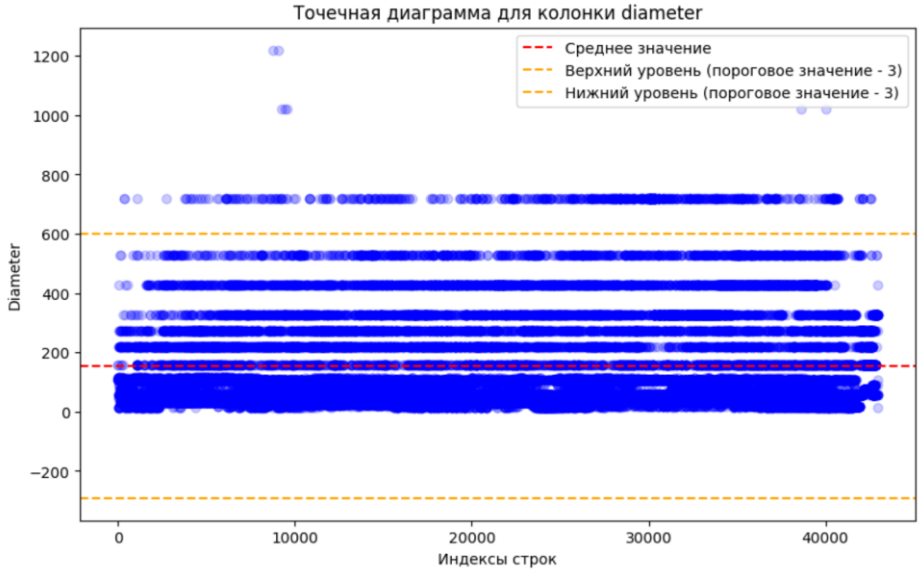

Figure 23 Точечная диаграмма для колонки "diameter"

Точечная диаграмма для переменной диаметр показала, что:

- Значения диаметра достаточно разбросаны вокруг среднего значения. Это свидетельствует о большой вариабельности диаметра в данных;
- На графике присутствуют выбросы, особенно в верхней части диапазона (значения диаметра выше 1000);
- На графике обозначены верхний и нижний пороговые значения, равные среднему значению плюс-минус 3 единицы. Это может использоваться для выделения значений, которые отклоняются от среднего более чем на 3 единицы;

- Распределение диаметра, скорее всего, скошено вправо. Это означает, что большая часть значений сосредоточена в левой части графика (ниже среднего), а хвост распределения тянется вправо (к более высоким значениям диаметра);
- На графике сложно определить точное количество мод (пиков), но можно предположить, что распределение может быть многомодальным, то есть иметь несколько пиков. Это может указывать на наличие нескольких подгрупп данных с различными значениями диаметра.

Идентификация выбросов по методу z-оценок

Проведены z-оценки для переменных temperature, diameter для выявления аномальных значений.

Элемент xi считается выбросом если величина

$$z_i = \frac{x_i - \bar{x}}{s}$$

либо меньше -3, либо больше 3 (или устанавливаем собственные пороговые значения при необходимости).

Переменная	**Количество аномальных значений**
temperature	125 аномальных значений при пороге 2.5
diameter	735 аномальных значений при пороге 3
temperature и diameter (расчет одновременно для обеих переменных)	По обеим колонкам 843 аномальных значения при пороге temperature = 2.5 и при пороге diameter = 3

Тест хи-квадрат для категориальных признаков

Тест хи-квадрат является статистическим инструментом, который используется для изучения взаимосвязи между категориальными переменными[11]. Ниже приведены результаты анализа:

Признак	**Статистика хи-квадрат**	**Уровень доверия**	**Проверка гипотезы**	**Характер взаимосвязи**
diameter	6228.80	~ 0.0000	Отвергаем основную гипотезу (H0), есть взаимосвязь с фактом брака (defect)	Слабая взаимосвязь
steel_grade	4127.23	~ 0.0000	Отвергаем основную гипотезу (H0), есть взаимосвязь с фактом брака (defect)	Слабая взаимосвязь
welding_method	1315.26	~ 0.0000	Отвергаем основную гипотезу (H0), есть взаимосвязь с фактом брака (defect)	Слабая взаимосвязь

[11] Квадратный тест: Изучение категориальных переменных в двух хвостовых тестах [Электронный ресурс] URL: https://fastercapital.com/ru/content/%

welding_materials	4766.44	~ 0.0000	Отвергаем основную гипотезу (H0), есть взаимосвязь с фактом брака (defect)	Слабая взаимосвязь

По результатам оценки с помощью теста хи-квадрат была выявлено наличие слабой взаимосвязи между каждой из исследуемых переменных с переменной интереса (зависимой переменной).

Коэффициент корреляции Пирсона для количественных признаков

Был оценен коэффициент корреляции для выявления характера взаимосвязи переменной “temperature” с зависимой переменной “defect”; была выявлена слабая положительная взаимосвязь:

Признак	**Коэффициент корреляции Пирсона**	**Уровень доверия**	**Проверка гипотезы**	**Характер взаимосвязи**
temperature	0.13	~ 0.00	Отвергаем основную гипотезу (H0), есть взаимосвязь с фактом брака (defect)	Слабая положительная взаимосвязь

Таким образом, возвращаясь к гипотезе исследования:

Гипотеза	**Уровень**	**Результат**	**Результат**

	значимости	**тестирования**	
H0: Положительная/отрицательная взаимосвязь между предоставленными переменными интереса (Xn: temperature, diameter, steel_grade, welding_method, welding_materials) и фактом отсутствия / наличия брака (defect) отсутствует. H1: Существует положительная/отрицательная взаимосвязь между предоставленными переменными интереса (Xn: temperature, diameter, steel_grade, welding_method, welding_materials) и фактом отсутствия / наличия брака	0,01	Отвергнута	Существует статистически значимая слабая положительная взаимосвязь

(defect).			

Корреляционная матрица

В последнюю очередь была построена корреляционная матрица для потенциальных корректировок в наборе данных в случае релевантной мультиколлинеарности; выявленная мультиколлинеарность между переменными может оказывать смещающие влияние на результаты регрессионной модели, однако, в случае удаления отдельных переменных модель будет неполной, было принято решение этим пренебречь.

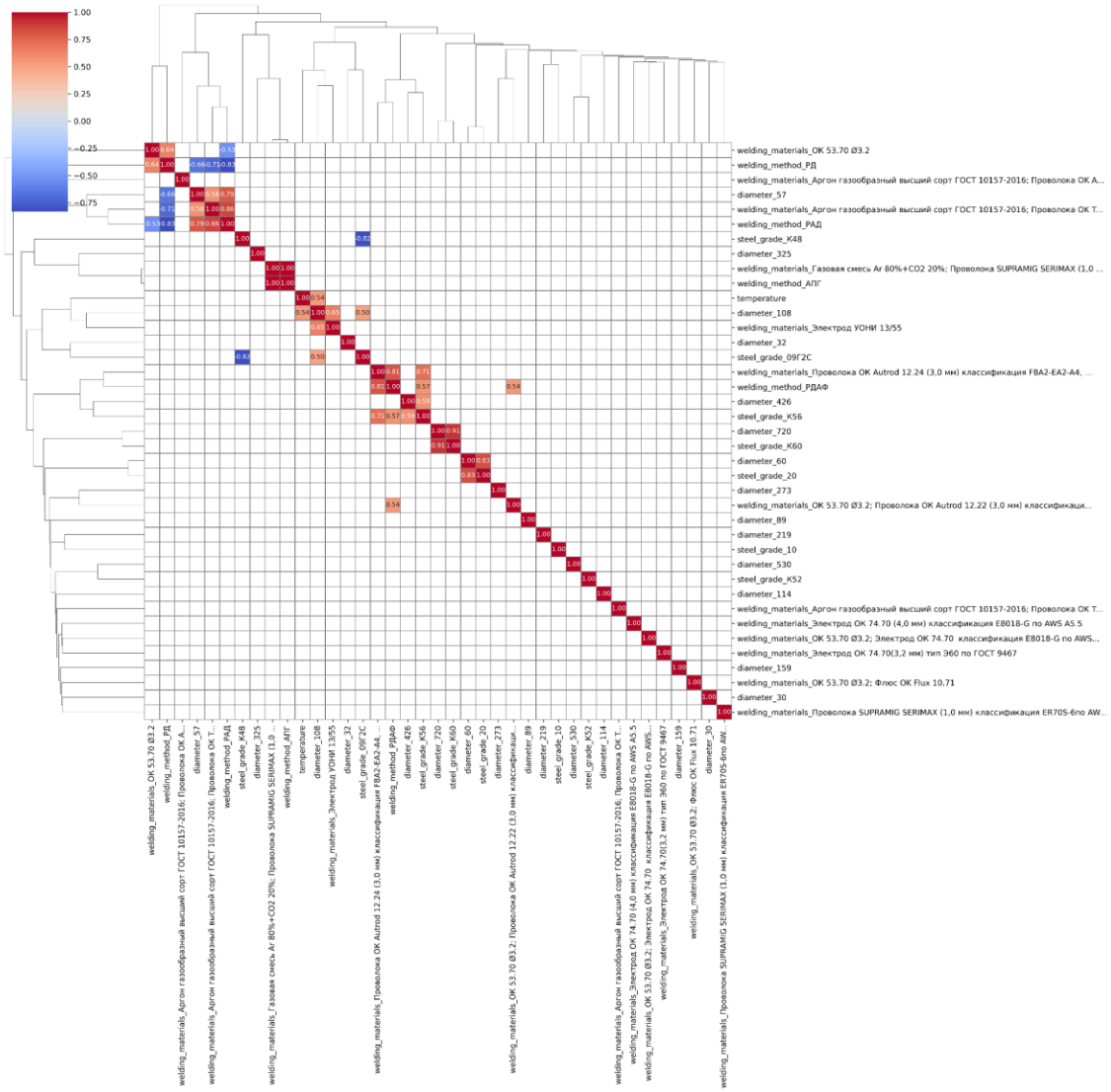

Figure 24 Корреляционная матрица

3.3. Методология исследования

Предварительная обработка категориальных переменных

Для начала категориальные переменные были преобразованы в удобный для проведения дальнейшего анализа вид:

```
categorical_features = ["diameter", "steel_grade", "welding_materials", "welding_method"]

df_one_hot = pd.get_dummies(df, columns=categorical_features, dtype="int")

independent = df_one_hot.drop(columns=["defect"])

dependent = df_one_hot["defect"]
```

Функция pd.get_dummies() выполняет кодирование категориальных признаков методом One-Hot Encoding[12]. Это означает, что для каждого значения категориального признака создаётся отдельный столбец, который принимает значение 1, если признак соответствует данному значению, и 0 в противном случае.

Макеты спецификаций моделей

К итоговой модели были предъявлены следующие требования: “Регрессионная модель должна представлять из себя модель, в которую включены исключительно переменные, коэффициенты при которых были оценены в рамках стандартных процедур проверки как статистически значимые. Переменные, в рамках которых гипотезы о статистической значимости которых были отвергнуты, должны быть исключены из модели”.

В предыдущей главе было указано, что каждый из рассматриваемых регрессоров был оценен как статистически значимый, соответветственно, имелась возможность выбрать различные спецификации модели для анализа в зависимости от метрик качества оценивания модели, а также исследовательских приоритетов.

Таким образом, исследование базируется на следующем бинарном многофакторном регрессионном уравнении с зависимой переменной, отвечающей за факт наличия или отсутствия бракованной детали по результатам трубосварочной операции.

Уравнение с зависимой переменной defect имеет следующий вид:

[12] Как получить полезную информацию из своих категориальных признаков? [Электронный ресурс] URL: https://habr.com/ru/companies/karuna/articles/769366/

$$defect = \beta_0 + \beta_1 temperature_{it} + \beta_2 diameter_{it} + \beta_3 steel\_grade_{it} + \beta_4 welding\_method_{it} + \beta_5 welding\_materials_{it} + \varepsilon_{it}$$

где i = 1, …, N, t = дд.мм.2023, дд.мм.2024

- В данном уравнении $defect_{it}$ выступает в качестве зависимой переменной и является переменной, отвечающей за факт наличия или отсутствия бракованной детали по результатам трубосварочной операции i в момент времени t;
- temperature - температура рабочей среды, при которой была совершена сварочная операция i в момент времени t, выраженная в градусах Цельсия;
- Вектор $diameter_{it}$ состоит из переменных, описывающих наружный диаметр элемента, выраженный в миллиметрах (мм), в рамках трубосварочной операции i в момент времени t;
- Вектор $steel\_grade_{it}$ состоит из переменных, описывающих марку стали элемента в рамках трубосварочной операции i в момент времени t;
- Вектор $welding\_method_{it}$ состоит из переменных, описывающих используемый при сварочной операции способ сварки в рамках трубосварочной операции i в момент времени t;
- Вектор $welding\_materials_{it}$ состоит из переменных, описывающих используемые при сварочной операции сварочные материалы в рамках трубосварочной операции i в момент времени t;
- Коэффициенты $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ являются неизвестными скалярными величинами при компонентах бинарного регрессионного многофакторного уравнения;
- ε_{it} – случайная величина, характеризующая ошибку аппроксимации.

3.4. Регрессионный анализ

Следующим шагом стало построение регрессионных моделей с помощью аналитического инструментария Python. Использованные библиотеки:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import statsmodels.api as sm
```

```
import matplotlib.pyplot as plt
```

```
coefficients = model.coef_[0]

logit_model = sm.Logit(y_train, X_train)
result = logit_model.fit(method='bfgs')
# Экспоненцирование коэффициентов для получения отношения шансов
odds_ratios = np.exp(coefficients)
print('Odds Ratios:')
print(odds_ratios)
```

```
df_test = df.loc[X_test.index]
preds_odds = model.predict_proba(X_test)
defect_pred = model.predict(X_test)
assurance_pred = np.max(preds_odds, axis=1)
df_test["defect_pred"] = defect_pred
df_test["assurance_pred"] = assurance_pred
df_test
```

3.5. Машинное обучение. Решение задачи классификации

Итак, модель логистической регрессии для задачи машинного обучения была задана с помощью следующей конструкции:

```
X_train, X_test, y_train, y_test = train_test_split(independent, dependent, test_size=0.3,
random_state=42)
if weight_balanced:
    model = LogisticRegression(class_weight="balanced")
else:
    model = LogisticRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

Ниже дано описание элементов в рамках конструкции.

```
X_train, X_test, y_train, y_test = train_test_split(independent, dependent, test_size=0.3,
random_state=42)
```

- Данные разделяются на обучающую (Обучающая выборка, тренировочная, англ.: training sample - выборка, по которой производится настройка, то есть оптимизация параметров модели зависимости. Если модель зависимости построена по обучающей выборке X^m, то оценка качества этой модели, сделанная по той же выборке X^m оказывается, как правило, оптимистически смещенной. Это нежелательное явление называют переобучением. На практике оно встречается очень часто. Хорошую эмпирическую оценку качества построенной модели дает её проверка на независимых данных, которые не использовались для обучения[13]) и тестовую (тестовая или контрольная выборка, англ.: test sample - выборка, по которой оценивается качество построенной модели. Если обучающая и тестовая выборки независимы, то оценка, сделанная по тестовой выборке, является несмещенной) выборки с использованием функции train_test_split;
- independent - это набор признаков (независимые переменные), а dependent - зависимая переменная, значения которой нужно предсказать;
- test_size = 0.3 указывает, что 30% данных будут отведены для тестирования, а 70% для обучения;
- random_state = 42 фиксирует случайное разделение, чтобы каждый раз при запуске кода данные делились одинаково.

```
if weight_balanced:
```

Эта строка проверяет условие weight_balanced. Если оно истинно, то используется модель с параметром class_weight="balanced", который автоматически балансирует веса классов (это может быть полезным, если у нас есть дисбаланс между классами). Например, если один класс встречается значительно реже, чем другой, модель будет корректировать веса, чтобы не обучиться на более часто встречающемся классе; Если weight_balanced — ложь, используется стандартная логистическая регрессия без балансировки весов.

```
model = LogisticRegression(class_weight="balanced")
```

[13] Выборка. Обучающая и тестовая выборка [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B0%D1%8E%D1%89%D0%B0%D1%8F_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0

Если условие weight_balanced истинно, создается модель логистической регрессии с параметром class_weight="balanced". Это означает, что модель будет автоматически учитывать дисбаланс классов при обучении.

```
model = LogisticRegression()
```

Если weight_balanced — ложь, создаётся стандартная модель логистической регрессии без балансировки классов.

```
model.fit(X_train, y_train)
```

Модель обучается на тренировочных данных. Функция fit принимает в качестве аргументов тренировочные признаки (X_train) и целевые значения (y_train); Таким образом, модель обучается находить зависимости между признаками и целевыми переменными.

```
y_pred = model.predict(X_test)
```

Модель предсказывает значения зависимой переменной для тестовых данных. Функция predict применяет обученную модель к тестовому набору признаков (X_test) и возвращает предсказания (в y_pred). Эти предсказания можно сравнить с реальными значениями y_test для оценки качества модели.

Метрики и методы оценивания модели машинного обучения

Для того, чтобы оценить, насколько качественной получилось аппроксимация зависимости между парами (пара есть признаковое описание и целевая переменная) по данным, доступным нам для анализа, были применены следующие метрики качества оценивания модели логистической регрессии (качество модели было оценено на тестовой выборке):

```
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
```

Перейдем к их описанию и анализу.

Model Accuracy (качество модели, агрегированное по всем классам)

Accuracy - это доля объектов, для которых мы правильно предсказали класс[14].

Необходимо иметь в виду, что для данной метрики есть несколько недостатков, основной из которых - метрика не учитывает дисбаланс классов (если в выборке есть сильный “перевес” одной категории над другой); на практике такая модель будет предсказывать значительную долю наблюдений (иметь высокую accuracy) по менее представленной категории в выборке. Помимо этого, не учитывается “цена ошибки” предсказания неверного класса (неверно предсказанная категория брака имеет меньшую значимость, чем неверно предсказанная категория “годной” детали).

Показатель в рамках нашей модели:

Accuracy	~ 0.80 (0.796272374976933)

Очевидно, что сразу же можно рассчитать **долю ошибочных классификаций (error rate)**:

$$Error\ Rate\ =\ 1\ -\ Accuracy$$

$$Error\ Rate\ =\ 1\ -\ 0.8\ =\ 0.2$$

Показатель Accuracy находится на приемлемом уровне; однако, учитывая факт того, что мы имеем значительный перевес по годным элементам (1 в наблюдении; то есть, классификатор, предсказывающий отсутствие брака, имеет достаточно высокую accuracy просто потому, что бракованных деталей в выборке намного меньше), стоит обратить внимание на прочие метрики модели.

Матрица ошибок (Confusion Matrix)

Внешнее представление матрицы в статистическом пакете для Python scikit-learn[15]:

[14] Учебник по машинному обучению от ШАД. Метрики классификации и регрессии [Электронный ресурс] URL: https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii

[15] Understanding the Confusion Matrix from Scikit learn [Электронный ресурс] URL: https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79

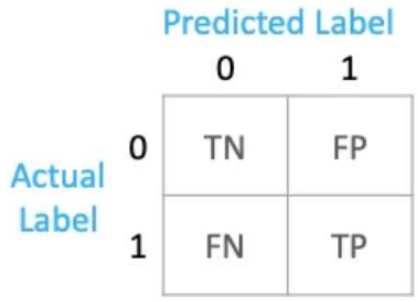

Figure 25 Что подразумевают под собой квадранты в матрице ошибок в пакете Scikit learn

Заметим, что для каждого объекта в выборке возможно 4 ситуации, которые и описывает матрица ошибок:

- Мы предсказали положительную метку (значение зависимой переменной) и угадали. Будет относить такие объекты к true positive (TP) группе. True — потому что предсказали мы правильно, а positive — потому что предсказали положительную метку;
- Мы предсказали положительную метку, но ошиблись в своём предсказании — false positive (FP). False, потому что предсказание было неправильным;
- Мы предсказали отрицательную метку и угадали — true negative (TN);
- И наконец, мы предсказали отрицательную метку, но ошиблись — false negative (FN).

Матрица ошибок (Confusion Matrix) для нашей модели:

	Предсказанное отрицательное значение (0)	**Предсказанное положительное значение (1)**
На самом деле отрицательное значение (0)	509 **TN (True Negative)** (количество верно предсказанных дефектов)	737 **FP (False Positive)**
На самом деле положительное значение (1)	367 **FN (False Negative)** (количество предсказанных	3806 **TP (True Positive)**

	ложноотрицательных дефектов)	

Говоря о смысле значений в рамках каждого квадранта в данном конкретном случае, можно сказать, что:

- **509** случаев, где модель правильно предсказала наличие дефектов (предсказанное значение = 0), и на самом деле дефекты были;
- **737** случаев, где модель предсказала отсутствие дефектов (предсказанное значение = 1), но на самом деле дефекты были;
- **367** случаев, где модель предсказала дефекты (предсказанное значение = 0), хотя дефекты на самом деле отсутствовали;
- **3806** случаев, где модель правильно предсказала отсутствие дефектов (предсказанное значение = 1), и дефекты действительно отсутствовали.

На основании данных результатов можно оценить такие метрики модели, как точность (precision), полнота (recall) и F1-мера.

Классификационный отчет (Classification Report)

В данном отчете представлены основные метрики в рамках модели.

Support - это количество реальных наблюдений в рамках класса.

Macro Avg - макросреднее (среднеарифметическое) для показателей в столбце.

Weighted Avg - взвешенное среднее для показателей в столбце.

Класс	***Precision***	***Recall***	***F1-Score***	***Support***
0	0.58	0.41	0.48	1246
1	0.84	0.91	0.87	4173
Accuracy			0.80	5419
Macro Avg	0.71	0.66	0.68	5419
Weighted Avg	0.78	0.80	0.78	5419

Рассмотрим метрики Precision, Recall, F1-Score из таблицы более подробно.

Точность (precision)

Если мы рассмотрим долю правильно предсказанных положительных объектов среди всех объектов, предсказанных положительным классом, то мы получим метрику, которая называется точностью (precision). Чем меньше ложноположительных срабатываний будет допускать модель, тем больше будет её Precision:

$$Precision = \frac{TP}{TP + FP} = \frac{3806}{3806 + 737} \approx 0{,}84$$

То есть, нашем случае она равна ~ 84% для годной детали.

Соответственно, Precision для бракованной детали ~58% (аналогично рассчитываем долю предсказанных негативных объектов среди всех объектов, предсказанных отрицательным классом), что является относительно приемлемым результатом.

Полнота (recall)

Если же мы рассмотрим долю правильно найденных положительных объектов среди всех объектов положительного класса, то мы получим метрику, которая называется полнотой (recall). Чем меньше ложно отрицательных срабатываний, тем выше recall модели:

$$Recall = \frac{TP}{TP + FN} = \frac{3806}{3806 + 367} \approx 0{,}91$$

В нашем случае она равна ~ 91% для годной детали (то есть, нам удалось выявить 91% годных деталей с помощью нашей модели); соответственно, Recall для бракованного элемента равен ~41%, что является относительно приемлемым результатом.

F1-мера (F1-score, F1-measure)

В случае пары Precision-Recall существует популярный способ скомпоновать их в одну метрику - взять их среднее гармоническое. Данный показатель эффективности исторически носит название F1-меры (F1-measure).

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0{,}84 * 0{,}91}{0{,}84 + 0{,}91} \approx 0{,}87$$

В нашем случае она равна ~ 87% для годной детали, что является приемлемым результатом.

Для бракованной детали метрика составляет 48%.

Таким образом, предсказательные метрики модели держатся на приемлемом уровне; модель дает относительно точные предсказательные оценки для определения того, будет ли элемент относится к категории “годен”; тем не менее, с точки зрения предсказания того, будет ли деталь бракованная или нет, модель справляется значительно хуже. Потенциально, можно сбалансировать модель в сторону предсказания брака в ущерб предсказательной силе того, будет ли элемент годен; данные доработки уже можно внедрять по мере определения нужд производства и бизнеса как такового.

Были также вычислены шансы получения бракованной и годной детали, где preds_odds = model.predict_proba(X_test) использует модель для предсказания вероятностей каждого класса для каждого объекта в тестовом наборе данных. Результатом был массив вероятностей, где каждая строка представляет вероятности для разных классов для соответствующего объекта.

```
preds_odds = model.predict_proba(X_test)
defect_pred = model.predict(X_test)
assurance_pred = np.max(preds_odds, axis=1)
df_test["defect_pred"] = defect_pred
df_test["assurance_pred"] = assurance_pred
```

Результаты были усреднены, после чего мы получили следующую таблицу:

Defect Prediction	**Assurance Prediction**
0	0.555690
1	0.840367

- 0.555690: для объекта с индексом 0 (брак) уверенность модели в предсказанном классе составляет 0.555690. Это означает, что модель оценивает вероятность

принадлежности этого объекта к предсказанному классу (бракованным элементам) как 55.57%.

- 0.840367: для объекта с индексом 1 (небракованный элемент) уверенность модели в предсказанном классе составляет 0.840367. Это означает, что модель оценивает вероятность принадлежности этого объекта к предсказанному классу (небракованным элементам) как 84.04%.

ГЛАВА 4. ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ ИССЛЕДОВАНИЯ. ПРЕДЛАГАЕМЫЕ ОРГАНИЗАЦИОННЫЕ РЕШЕНИЯ

Проведенный регрессионный анализ с помощью алгоритма логистической регрессии позволил определить предельные эффекты каждой переменной и оценить влияние различных факторов сварочного производства на вероятность получения брака. Ниже приведены показатели влияния каждой из переменных на вероятность брака выраженные в процентах (0,01 считается за 1%). То есть наличие каждого из факторов может как повышать вероятность брака на заданное количество процентов (если число отрицательное), так и снижать его вероятность (если число положительное).

Наименование переменной	**Влияние на вероятность брака**
temperature	-0,00022261870588966365
diameter_30	0.026156183971745378
diameter_32	0.5396085742136391
diameter_57	0.38719930342357406
diameter_60	0.09974817361975684
diameter_89	0.055671463321642224
diameter_108	-0.16816769286215014
diameter_114	-0.04332688938554826
diameter_159	0.09097243485426906
diameter_219	-0.1234129794782614
diameter_273	-0.22153702735005007
diameter_325	-0.2818190420154787
diameter_426	-0.13459359070252866
diameter_530	-0.171978565632896
diameter_720	0.050303741577574786
steel_grade_10	0.008020454664449532
steel_grade_20	0.10300775168939989
steel_grade_09Г2С	0.13161597243581957
steel_grade_К48	-0.1792471375846129
steel_grade_К52	0.010652836469251914
steel_grade_К56	0.16310187040544957
steel_grade_К60	-0.13232766052457504
welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Autrod 13.23 классификация ER80S-Ni1по AWS A5.28	0.131228650797549
welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 12.64(2,4 мм) классификация ER70S-6по AWS A5.18	-0.38990444376310057

welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 13.23(2,4 мм) классификация ER80S-Ni1по AWS A5.28	0.11473984170100662
welding_materials_Газовая смесь Ar 80%+CO2 20%; Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18	0.050185999429851046
welding_materials_ОК 53.70 Ø3.2	-0.19613895466353365
welding_materials_ОК 53.70 Ø3.2; Проволока ОК Autrod 12.22 (3,0 мм) классификация E7A4-EMпо AWS A5.23; Флюс ОК Flux 10.71	0.2745657151541723
welding_materials_ОК 53.70 Ø3.2; Флюс ОК Flux 10.71	0.09364995124341693
welding_materials_ОК 53.70 Ø3.2; Электрод ОК 74.70 классификация E8018-G по AWS A5.5	0.03476307520301639
welding_materials_Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6по AWS A5.18	-0.11386937053699976
welding_materials_Проволока ОК Autrod 12.24 (3,0 мм) классификация F8A2-EA2-A4, F7P0-EA2-A4по AWS A5.23; Флюс ОК Flux 10.71; Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	-0.18786551330467272
welding_materials_Электрод ОК 74.70 (4,0 мм) классификация E8018-G по AWS A5.5	-0.22923998740319412
welding_materials_Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467	0.03411465173608964
welding_materials_Электрод УОНИ 13/55	0.48859447196139355
welding_method_АПГ	-0.06368337110715269
welding_method_РАД	-0.14393595126447098
welding_method_РД	0.13209325683388515
welding_method_РДАФ	0.18035015309294064

ЗАКЛЮЧЕНИЕ

На основе проведённого анализа можно сделать следующие выводы:

1. **Влияние температуры (temperature):**

   - предельный эффект для переменной temperature составляет -0,00022261870588966365. Отрицательное значение указывает на то, что увеличение температуры может повышать вероятность получения брака на 0,000223%. Однако этот эффект очень мал и может быть незначительным по сравнению с другими факторами.

2. **Влияние диаметра трубопровода:**

   - **diameter_30**: предельный эффект равен 0,026156183971745378. Это означает, что использование электродов диаметром 30 мм может снизить вероятность брака на 2,62%.
   - **diameter_32**: предельный эффект составляет 0,5396085742136391. Использование электродов диаметром 32 мм может значительно снизить вероятность брака — на 53,96%.
   - **diameter_57**: предельный эффект 0,38719930342357406. Применение электродов диаметром 57 мм способно уменьшить вероятность брака на 38,72%.
   - **diameter_60**: предельный эффект 0,09974817361975684. Использование электродов диаметром 60 мм может уменьшить вероятность брака примерно на 10%.
   - **diameter_89**: предельный эффект 0,055671463321642224. Применение электродов диаметром 89 мм потенциально снижает вероятность брака на 5,57%.
   - **diameter_108**: предельный эффект -0,16816769286215014. Использование электродов диаметром 108 мм может повысить вероятность брака на 16,82%
   - **diameter_114**: предельный эффект составляет -0,04332688938554826. Использование трубопроводов диаметром 114 мм может повысить вероятность брака примерно на 4,33%.

- **diameter_159**: предельный эффект равен 0,09097243485426906. Применение трубопроводов диаметром 159 мм потенциально снижает вероятность брака на 9,10%.
- **diameter_219**: предельный эффект -0,1234129794782614. Использование трубопроводов диаметром 219 мм может повысить вероятность брака приблизительно на 12,34%.
- **diameter_273**: предельный эффект -0,22153702735005007. Применение трубопроводов диаметром 273 мм может повысить вероятность возникновения дефектов на 22,15%.
- **diameter_325**: предельный эффект -0,2818190420154787. Использование трубопроводов диаметром 325 мм может повысить вероятность брака примерно на 28,18%.
- **diameter_426**: предельный эффект -0,13459359070252866. Применение трубопроводов диаметром 426 мм может повысить вероятность дефектов приблизительно на 13,46%.
- **diameter_530**: предельный эффект -0,171978565632896. Использование трубопроводов диаметром 530 мм может повысить вероятность получения брака примерно на 17,20%.
- **diameter_720**: предельный эффект 0,050303741577574786. Применение трубопроводов диаметром 720 мм потенциально уменьшает вероятность дефектов на 5,03%.

3. **Влияние марки стали на вероятность получения брака:**

- **steel_grade_10**: предельный эффект составляет 0,008020454664449532. Использование стали марки 10 может снизить вероятность возникновения брака примерно на 0,80%.
- **steel_grade_20**: предельный эффект равен 0,10300775168939989. Применение стали марки 20 потенциально снижает вероятность брака на 10,30%.

- **steel_grade_09Г2С**: предельный эффект 0,13161597243581957. Использование стали 09Г2С может уменьшить вероятность брака приблизительно на 13,16%.
- **steel_grade_К48**: предельный эффект -0,1792471375846129. Применение стали К48 может повысить вероятность дефектов примерно на 17,92%.
- **steel_grade_К52**: предельный эффект 0,010652836469251914. Использование стали К52 может снизить вероятность получения брака примерно на 1,07%.
- **steel_grade_К56**: предельный эффект 0,16310187040544957. Применение стали К56 может уменьшить вероятность возникновения дефектов приблизительно на 16,31%.
- **steel_grade_К60**: предельный эффект -0,13232766052457504. Использование стали К60 может повысить вероятность брака примерно на 13,23%.

4. **Влияние сварочных материалов на вероятность получения брака:**

- **welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Autrod 13.23 классификация ER80S-Ni1 по AWS A5.28**: предельный эффект составляет 0,131228650797549. Использование этих сварочных материалов может снизить вероятность возникновения брака примерно на 13,12%.
- **welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 12.64 (2,4 мм) классификация ER70S-6 по AWS A5.18**: предельный эффект равен -0,38990444376310057. Применение этих сварочных материалов потенциально повышает вероятность дефектов на 38,99%.
- **welding_materials_Аргон газообразный высший сорт ГОСТ 10157-2016; Проволока ОК Tigrod 13.23 (2,4 мм) классификация ER80S-Ni1 по AWS A5.28**: предельный эффект составляет 0,11473984170100662. Использование этих сварочных материалов может повысить вероятность возникновения брака примерно на 11,47%.

- **welding_materials_Газовая смесь Ar 80%+CO_2 20%; Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6 по AWS A5.18**: предельный эффект равен 0,050185999429851046. Применение этих сварочных материалов потенциально уменьшает вероятность дефектов на 5,02%.
- **welding_materials_ОК 53.70 Ø3.2**: предельный эффект составляет -0,19613895466353365. Использование этих сварочных материалов может повысить вероятность возникновения брака примерно на 19,61%.
- **welding_materials_ОК 53.70 Ø3.2; Проволока ОК Autrod 12.22 (3,0 мм) классификация E7A4-EM по AWS A5.23; Флюс OK Flux 10.71**: предельный эффект равен 0,2745657151541723. Применение этих сварочных материалов потенциально уменьшает вероятность дефектов на 27,46%.
- **welding_materials_ОК 53.70 Ø3.2; Флюс OK Flux 10.71**: предельный эффект 0,09364995124341693. Использование данных сварочных материалов может уменьшить вероятность брака приблизительно на 9,36%.
- **welding_materials_ОК 53.70 Ø3.2; Электрод ОК 74.70 классификация E8018-G по AWS A5.5**: предельный эффект составляет 0,03476307520301639. Использование этих сварочных материалов может снизить вероятность возникновения брака примерно на 3,48%.
- **welding_materials_Проволока SUPRAMIG SERIMAX (1,0 мм) классификация ER70S-6 по AWS A5.18**: предельный эффект равен -0,11386937053699976. Применение этих сварочных материалов потенциально повышает вероятность дефектов на 11,39%.
- **welding_materials_Проволока ОК Autrod 12.24 (3,0 мм) классификация F8A2-EA2-A4, F7P0-EA2-A4 по AWS A5.23; Флюс OK Flux 10.71; Электрод ОК 74.70(3,2 мм) тип Э60 по ГОСТ 9467**: предельный эффект -0,18786551330467272. Использование данных сварочных материалов может повысить вероятность брака приблизительно на 18,79%.

- Использование **сварочных материалов ОК 53.70 Ø3.2 и флюса OK Flux 10.71** увеличивает вероятность получения качественного изделия на 0,09%.

- Применение **электродов ОК 74.70 (3,2 мм) типа Э60 по ГОСТ 9467** также положительно влияет на качество продукции, снижая вероятность брака на 0,03%.

- **Проволока SUPRAMIG SERIMAX (1,0 мм) классификации ER70S-6по AWS A5.18** уменьшает риск возникновения дефектов на 0,11%.

- **Использование проволоки ОК Autrod 12.24 (3,0 мм), флюса ОК Flux 10.71 и электрода ОК 74.70(3,2 мм)** снижает вероятность брака на 0,19%.

- **Электрод ОК 74.70 (4,0 мм) классификация E8018-G по AWS A5.5** уменьшает вероятность возникновения брака на 0,23%.

- **Электрод УОНИ 13/55** повышает вероятность получения качественного результата на 49%.

5. **Влияние методов сварки на вероятность получения брака:**

- Применение метода автоматической дуговой сварки плавящимся электродом АПГ повышает вероятность получения брака на 6%.

- Применение метода сварки РАД повышает вероятность брака на 14%.

- Применение метода сварки РДАФ снижает вероятность брака на 18%.

- Применение метода сварки РД может снизить вероятность получения брака на 13%.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Stock, James H, and Mark W. Watson. Introduction to Econometrics [Электронный ресурс] URL: https://www.sea-stat.com/wp-content/uploads/2020/08/James-H.-Stock-Mark-W.-Watson-Introduction-to-Econometrics-Global-Edition-Pearson-Education-Limited-2020.pdf
2. What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices [Электронный ресурс] URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/
3. Учебники Экономического факультета МГУ. Дружелюбная эконометрика [Электронный ресурс] URL: https://books.econ.msu.ru/Introduction-to-Econometrics/
4. Критерий Шапиро-Уилка [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A8%D0%B0%D0%BF%D0%B8%D1%80%D0%BE-%D0%A3%D0%B8%D0%BB%D0%BA%D0%B0
5. Критерий Колмогорова-Смирнова [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%9A%D0%BE%D0%BB%D0%BC%D0%BE%D0%B3%D0%BE%D1%80%D0%BE%D0%B2%D0%B0-%D0%A1%D0%BC%D0%B8%D1%80%D0%BD%D0%BE%D0%B2%D0%B0
6. Критерий Андерсона-Дарлинга [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%90%D0%BD%D0%B4%D0%B5%D1%80%D1%81%D0%BE%D0%BD%D0%B0-%D0%94%D0%B0%D1%80%D0%BB%D0%B8%D0%BD%D0%B3%D0%B0
7. Как выполнить тест Андерсона-Дарлинга в Python [Электронный ресурс] URL: https://www.codecamp.ru/blog/anderson-darling-test-python/
8. Асимметрия и эксцесс эмпирического распределения [Электронный ресурс] URL: http://mathprofi.ru/asimmetriya_i_excess.html
9. ГРАФИКИ КВАНТИЛЬ-КВАНТИЛЬ (Q – Q) ДЛЯ ГИДРОЛОГО-ГИДРОЛОГИЧЕСКИХ ДАННЫХ МОЛДОВЫ [Электронный ресурс] URL: https://ibn.idsi.md/sites/default/files/imag_file/p-75-78_3.pdf

10. Диаграмма размаха ("ящик с усами") [Электронный ресурс] URL: https://datavizcatalogue.com/RU/metody/diagramma_razmaha.html
11. Квадратный тест: Изучение категориальных переменных в двух хвостовых тестах [Электронный ресурс] URL: https://fastercapital.com/ru/content/%
12. Как получить полезную информацию из своих категориальных признаков? [Электронный ресурс] URL: https://habr.com/ru/companies/karuna/articles/769366/
13. Выборка. Обучающая и тестовая выборка [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B0%D1%8E%D1%89%D0%B0%D1%8F_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0
14. Учебник по машинному обучению от ШАД. Метрики классификации и регрессии [Электронный ресурс] URL: https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii
15. Understanding the Confusion Matrix from Scikit learn [Электронный ресурс] URL: https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79

Для написания исследования также были изучены следующие статьи, учебники:

1. Что такое ловушка с фиктивной переменной? (Определение и пример) [Электронный ресурс] URL: https://www.codecamp.ru/blog/dummy-variable-trap/
2. Логистическая регрессия и ROC-анализ — математический аппарат [Электронный ресурс] URL: https://loginom.ru/blog/logistic-regression-roc-auc#:~:text=%
3. Выбор уровня значимости при проверке статистических гипотез [Электронный ресурс] URL: https://www.statmethods.ru/stati/vybor-urovnya-znachimosti-pri-proverke-statisticheskikh-gipotez/
4. Предварительная обработка данных [Электронный ресурс] URL: https://habr.com/ru/articles/511132/
5. Обработка данных для машинного обучения [Электронный ресурс] URL: https://4brain.ru/aibasics/data.php
6. Категориальные признаки [Электронный ресурс] URL: https://habr.com/ru/articles/666234/
7. Прикладной анализ данных в социальных науках. Учебник от Yandex [Электронный ресурс] URL: https://education.yandex.ru/handbook/data-analysis

8. Logistic Regression Playlist by StatQuest [Электронный ресурс] URL: https://youtube.com/playlist?list=PLblh5JKOoLUKxzEP5HA2dLi7IJkHfXSe&si=hoUJPuhOjM3EBvRD
9. Odds and Log(Odds), Clearly Explained!!! By StatQuest [Электронный ресурс] URL: https://youtu.be/ARfXDSkQf1Y?si=1pG5FApzYdsDhcbh
10. Odds Ratios and Log(Odds Ratios), Clearly Explained!!! By StatQuest [Электронный ресурс] URL: https://youtu.be/8nm0G-1uJzA?si=TcLjQOTGQou9KObW
11. Эконометрика (Курс Открытое образование х ВШЭ) [Электронный ресурс] URL: https://apps.openedu.ru/learning/course/course-v1:hse+METRIX+2022/home
12. Логистическая регрессия, самое простое объяснение! От Start Career in Data Science [Электронный ресурс] URL: https://youtu.be/8sp5aqyH6Oc?si=Yv7Z9MnoXpD0xovf
13. Семинары Кислицына Д. В. (к.э.н., доцент департамента экономики ВШЭ) по эконометрике для Высшей Школы Менеджмента