МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science Pro»

«Построение прогнозной модели прочности бетона на сжатие по данным о компонентном составе смеси»

Слушатель Ялковский К. А.

Москва, 2025

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

СОДЕРЖАНИЕ

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

ВВЕДЕНИЕ

Актуальность:

Повышение точности прогнозирования прочности бетона позволяет оптимизировать расход материалов и обеспечить надежность строительных конструкций, что напрямую влияет на экономическую эффективность и безопасность в строительной отрасли.

Цель работы:

Разработать и сравнить точность моделей машинного обучения для прогнозирования прочности бетона на сжатие на основе данных компонентном составе смеси, интегрировать лучшую модель в приложение с графическим интерфейсом.

Задачи исследования:

1. Провести анализ и предобработку данных;
2. Реализовать и обучить различные модели машинного обучения;
3. Сравнить точность моделей и выбрать оптимальную;
4. Разработать приложение с графическим интерфейсом;
5. Протестировать работоспособность решения.

Исходные данные:

Набор данных взят из репозитория UCI Machine Learning Repository:

https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength

Первоначальный автор и донор набора данных:

Проф. И-Ченг Йе (I-Cheng Yeh)

Департамент информационного менеджмента

Университет Чунг-Хуа

Синьчжу, Тайвань 30067, Китайская Республика

E-mail: icyeh@chu.edu.tw

Тел.: 886-3-5186511

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

1. АНАЛИТИЧЕСКАЯ ЧАСТЬ

1.1. Постановка задачи

Формулировка задачи: Имея на входе данные о компонентах бетонной смеси (цемент, доменный шлак, зола-унос, вода, суперпластификатор, крупный и мелкий заполнитель) и возрасте твердения, необходимо спрогнозировать прочность бетона на сжатие в МПа.

Описание датасета: 1030 наблюдений, 8 входных переменных, 1 выходная переменная.

Таблица 1– Описание переменных

Переменная	**Тип данных**	**Единица измерения**	**Описание**
Цемент (компонент 1)	количественная	кг в $м^3$смеси	Входная переменная
Гранулированный доменный шлак (компонент 2)	количественная	кг в $м^3$смеси	Входная переменная
Зола-унос (компонент 3)	количественная	кг в $м^3$смеси	Входная переменная
Вода (компонент 4)	количественная	кг в $м^3$смеси	Входная переменная
Суперпластификатор (компонент 5)	количественная	кг в $м^3$смеси	Входная переменная
Крупный заполнитель (компонент 6)	количественная	кг в $м^3$смеси	Входная переменная
Мелкий заполнитель (компонент 7)	количественная	кг в $м^3$смеси	Входная переменная
Возраст	количественная	Дни	Входная переменная
Прочность бетона на сжатие	количественная	МПа	Целевая переменная

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Таблица 2 – Описательная статистика

Переменная	**Мин**	**Макс**	**Среднее**	**Медиана**	**Стандартное отклонение**
Цемент	102	540	281.2	272.9	104.5
Гранулированный доменный шлак	0	359.4	73.9	22	86.3
Зола-унос	0	200.1	52.2	0	64
Вода	121.8	247	181.6	185	21.4
Суперпластификатор	0	32.2	6.2	6.4	6
Крупный заполнитель	801	1145	972.9	968	77.8
Мелкий заполнитель	594	962.6	773.6	779.5	80.2
Возраст	1	365	45.7	28	63.2
Прочность бетона на сжатие	2.3	82.6	35.8	34.4	16.7

Характеристика набора данных: данные в сыром виде, пропуски отсутствуют, найдено и удалено 25 дубликатов.

1.2. Описание используемых методов

Линейная регрессия

Модель предсказывает целевую переменную как линейную комбинацию признаков.

Плюсы:

– Высокая интерпретируемость - легко понять, как каждый признак влияет на результат.

– Очень высокая скорость работы - обучение и предсказание происходят почти мгновенно.

– Не склонна к переобучению при небольшом количестве признаков.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Простота реализации и понимания.

Минусы:

– Строгие требования к данным: предполагает линейную зависимость, нормальное распределение ошибок и отсутствие сильной корреляции между признаками;

– Не может уловить сложные нелинейные закономерности в данных;

– Сильная чувствительность к выбросам в данных;

– Плохо работает, когда признаков больше, чем наблюдений.

Когда использовать:

– Как базовую модель для сравнения с более сложными алгоритмами.

– Когда критически важна интерпретируемость модели.

– В задачах, где есть явная линейная зависимость между признаками и целевой переменной.

– При серьезных ограничениях вычислительных ресурсов.

Главные предпосылки: наличие линейной зависимости между целевой переменной и признаками, отсутствие сильной мультиколлинеарности.

Lasso-регрессия

Это модификация линейной регрессии с L1-регуляризацией. Модель добавляет к функции потерь штраф, равный сумме абсолютных значений коэффициентов. Это заставляет модель не только минимизировать ошибку, но и уменьшать веса признаков.

Плюсы:

– Автоматический отбор признаков – неважные признаки получают нулевые веса;

– Борется с переобучением лучше обычной линейной регрессии;

– Улучшает обобщающую способность модели;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Сохраняет частичную интерпретируемость.

Минусы:

– При наличии сильно коррелированных признаков выбирает случайный один из них

– Может исключить полезные признаки, если параметр регуляризации слишком велик

– Медленнее обычной линейной регрессии

– Требует тщательного подбора гиперпараметра регуляризации

Когда использовать:

– Когда нужно сократить количество признаков в модели;

– При работе с данными, где много потенциально нерелевантных признаков;

– Когда нужен компромисс между интерпретируемостью и качеством;

– Для борьбы с мультиколлинеарностью.

Главные предпосылки: разреженность данных (много нулевых или малозначимых признаков), линейная зависимость. Менее чувствителен к корреляции признаков, чем Ridge, но все же страдает при сильной мультиколлинеарности.

ElasticNet

Комбинированный метод, объединяющий L1-регуляризацию (как в Lasso) и L2-регуляризацию (как в Ridge). Штрафная функция включает сумму абсолютных значений коэффициентов и сумму их квадратов.

Плюсы:

– Сочетает преимущества Lasso и Ridge-регрессии;

– Устойчив к сильной корреляции между признаками – не случайно выбирает один признак из группы;

– Эффективно отбирает признаки как Lasso, но более стабильно;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Лучшая предсказательная способность при мультиколлинеарности.

Минусы:

– Сложнее в настройке – требует подбора двух гиперпараметров вместо одного;

– Вычислительно более затратен чем Lasso или Ridge по отдельности;

– Медленнее сходится при оптимизации;

– Менее интерпретируем чем чистый Lasso.

Когда использовать:

– Когда в данных есть группы сильно коррелированных признаков;

– Когда количество признаков больше количества наблюдений;

– Как компромисс между отбором признаков и устойчивостью к мультиколлинеарности.

– Часто как основная линейная модель по умолчанию

Главные предпосылки: линейная зависимость, наличие групп коррелированных признаков.

Kernel Ridge Regression

Комбинация гребневой регрессии (Ridge) с ядерным методом. Сначала данные нелинейно преобразуются в пространство высшей размерности через ядро, затем в этом пространстве строится линейная регрессия с L2-регуляризацией.

Плюсы:

– Может моделировать сложные нелинейные зависимости;

– Регуляризация предотвращает переобучение в пространстве признаков;

– Теоретически обоснованный метод с закрытой формой решения;

– Гибкость за счет выбора разных ядер (RBF, полиномиальное);

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Минусы:

– Крайне медленное обучение – время растет кубически от числа объектов;

– Практически неприменим для наборов данных больше 10-20 тысяч наблюдений;

– Требует тщательного подбора ядра и гиперпараметров;

– Плохая интерпретируемость – работает как черный ящик.

Когда использовать:

– На небольших наборах данных (сотни, тысячи наблюдений);

– Для моделирования нелинейных зависимостей.

Главные предпосылки: небольшой объем данных, наличие нелинейных зависимостей, которые можно уловить выбранным ядром. Критически зависит от выбора гиперпараметров регуляризации и ядра.

Decision Tree для регрессии

Древовидная модель, которая рекурсивно разбивает данные на подгруппы по значениям признаков. В каждом листе дерева предсказанием является среднее значение целевой переменной объектов, попавших в этот лист.

Плюсы:

– Полная интерпретируемость – можно проследить весь путь предсказания;

– Не требует предобработки данных (нормализации, масштабирования);

– Может улавливать нелинейные зависимости и взаимодействия признаков;

– Работает с категориальными и числовыми признаками без преобразований.

Минусы:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Сильное переобучение – дерево может подстроиться под шум в данных;

– Неустойчивость – небольшие изменения данных могут сильно менять структуру дерева;

– Плохая обобщающая способность без ограничения глубины;

– Склонность к захвату выбросов;

– Не может предсказать значения вне диапазона обучения.

Когда использовать:

– Когда максимально важна интерпретируемость модели;

– Для быстрого прототипирования и анализа данных;

– В задачах, где важны бизнес-правила и логика принятия решений;

– Как базовая модель для ансамблевых методов (случайный лес, бустинг).

Главные предпосылки: наличие выраженных правил разбиения в данных, отсутствие строгих требований к стабильности предсказаний. Требует ограничения сложности.

Random Forest для регрессии

Ансамбль из множества деревьев решений, построенных на разных подвыборках. Итоговое предсказание – среднее предсказание всех деревьев.

Плюсы:

– Высокая точность по сравнению с одним деревом;

– Устойчивость к переобучению за счет усреднения;

– Меньшая чувствительность к шуму и выбросам;

– Возможность оценивать важность признаков;

– Работает с нелинейными зависимостями;

– Параллелизуется и хорошо масштабируется.

Минусы:

– Потеря интерпретируемости по сравнению с одним деревом;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Медленнее в обучении и предсказании чем одно дерево;
– Требует больше памяти;
– Не может экстраполировать за пределы обучающей выборки;
– Склонен к переобучению на зашумленных данных при глубоких деревьях.

Когда использовать:

– Когда нужна высокая точность без тонкой настройки;
– Для работы с разнородными данными;
– Когда важна устойчивость модели;
– Для оценки важности признаков;
– Как надежный метод по умолчанию для табличных данных.

Главные предпосылки: наличие достаточного количества данных, разнообразие признаков. Эффективен, когда отдельные деревья в ансамбле достаточно точны и разнообразны.

Gradient Boosting для регрессии

Последовательный ансамбль деревьев, где каждое следующее дерево учится предсказывать ошибки (остатки) предыдущих деревьев. Модель строится постепенно, минимизируя функцию потерь через градиентный спуск.

Плюсы:

– Очень высокая точность, часто лучшая среди методов машинного обучения;
– Гибкость – работает с разными функциями потерь;
– Автоматически учитывает взаимодействия признаков;
– Хорошо работает с нелинейными зависимостями;
– Может обрабатывать пропущенные значения.

Минусы:

– Склонность к переобучению без регуляризации;
– Долгое время обучения из-за последовательного подхода;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Чувствительность к гиперпараметрам и шуму в данных;
– Требует тщательной настройки;
– Вычислительно затратен.

Когда использовать:

– Когда нужна максимальная точность предсказаний;
– На структурированных табличных данных;
– Когда есть время и ресурсы для тонкой настройки.

Главные предпосылки: достаточное количество данных для последовательного обучения, правильный подбор скорости обучения и количества деревьев. Требует аккуратной регуляризации для избежания переобучения.

XGBoost (Extreme Gradient Boosting)

Это усовершенствованная реализация градиентного бустинга на решающих деревьях, которая сочетает высокую точность предсказаний с вычислительной эффективностью.

Основные принципы работы:

– Алгоритм строит последовательность деревьев, где каждое следующее дерево корректирует ошибки предыдущих;
– Использует вторые производные для более точной оптимизации функции потерь;
– Включает регуляризацию для контроля сложности модели;
– Эффективно обрабатывает пропущенные значения в данных;
– Поддерживает параллельные вычисления.

Преимущества:

– Высокая прогнозная точность на структурированных данных;
– Быстрая скорость работы благодаря оптимизациям;
– Встроенная защита от переобучения через регуляризацию;
– Автоматическая обработка пропущенных значений;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Возможность использования досрочной остановки.

– Гибкая система гиперпараметров для тонкой настройки.

Недостатки:

– Сложность интерпретации модели;

– Требует тщательного подбора параметров для достижения оптимальной производительности;

– Может переобучаться при неправильной настройке;

– Вычислительная сложность на больших объемах данных;

– Чувствительность к выбросам и шуму в данных.

Области применения:

– Задачи прогнозирования на табличных данных со сложными зависимостями;

– Продакшен-системы, где важны и точность, и скорость работы;

– Сценарии с достаточными вычислительными ресурсами для обучения;

– Критические факторы успеха:

– Достаточный объем данных для обучения;

– Правильный подбор скорости обучения и количества деревьев;

– Наличие сложных нелинейных взаимосвязей между признаками.

– LightGBM (Light Gradient Boosting Machine)

Высокоэффективная реализация градиентного бустинга, разработанная Microsoft с фокусом на скорость обучения и обработку больших объемов данных.

Основные технологические особенности:

– Использует Gradient-based One-Side Sampling (GOSS) для выборочного использования объектов;

– Применяет Exclusive Feature Bundling (EFB) для группировки разреженных признаков;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Работает по стратегии роста дерева leaf-wise, а не level-wise;

– Оптимизирован для работы с большими датасетами и высокоразмерными данными.

Преимущества:

– Значительно более высокая скорость обучения по сравнению с другими бустингами;

– Экономичное использование памяти благодаря оптимизированным структурам данных;

– Эффективная работа с категориальными признаками без предварительного кодирования;

– Поддержка параллельных и распределенных вычислений;

– Способность обрабатывать данные, не помещающиеся в оперативную память;

– Высокая прогнозная точность при правильной настройке.

Недостатки:

– Может переобучаться на небольших датасетах из-за leaf-wise роста;

– Требует тщательной настройки гиперпараметров для достижения оптимальной производительности;

– Менее стабильные результаты на маленьких выборках данных.

Области применения:

– Работа с очень большими наборами данных (миллионы строк);

– Задачи с высокоразмерными признаковыми пространствами;

– Системы реального времени, требующие быстрого переобучения;

– Промышленные приложения с ограниченными вычислительными ресурсами;

– Сценарии с преобладанием категориальных признаков.

– Критические факторы успеха:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

– Большой объем тренировочных данных для устойчивости leaf-wise роста;

– Правильная настройка параметров регуляризации;

– Адекватный подбор скорости обучения и количества деревьев;

– Учет специфики категориальных признаков в данных.

Таблица 3 – Сравнительная таблица методов для задачи регрессии

Метод	**Преимущества**	**Недостаки**	**Применимость**
Linear Regression	Интерпретируемость, Высокая скорость обучения, Способность к экстраполяции	Не может улавливать нелинейные зависимости, Чувствительна к масштабу данных	Отправная точка для решения задачи регрессии
Lasso	Обнуляет малозначимые признаки, Борется с переобучением	При наличии сильно коррелирующих признаков произвольно выбирает один из них	Разреженность данных (много нулевых или малозначимых признаков)
Elastic Net	Комбинирует L1 и L2 регуляризацию	Требует тщательной настройки гиперпараметров, Не улавливает нелинейные зависимости	Подходит для многомерных данных, Проводит отбор признаков и решает проблему мультиколлинеарности
Kernel Ridge	Позволяет описывать сложные нелинейные зависимости	Высокая вычислительная сложность	Нелинейные зависимости, небольшой набор данных

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Продолжение таблицы

Decision Tree	Позволяет моделировать нелинейные зависимости, Интерпретируемость, Не требует масштабирования данных	Склонна к переобучению, Нестабильность – небольшие изменения в данных могут привести к созданию совершенно другого дерева, Не может экстраполировать	Базовая модель для ансамблевых методов
Random Forest	Повышенная точность предсказания в сравнении с одним деревом, Устойчива к переобучению	Высокая вычислительная сложность, Меньшая интепретируемость	Наличие выраженных правил разбиения в данных, отсутствие строгих требований к стабильности предсказаний
Gradient Boosting	Высокая точность	Требует тщательной настройки гиперпараметров, Склонна к переобучению, Высокая вычислительная сложность	Достаточное количество данных для последовательного обучения; Правильный подбор скорости обучения и количества деревьев; Требует регуляризации для избежания переобучения.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Продолжение таблицы

XGBoost	Высокая точность Встроенная регуляризация; Автоматическая обработка пропущенных значений; Гибкая система гиперпараметров для тонкой настройки	Сложность интерпретации модели; Требует тщательного подбора параметров для достижения оптимальной производительности; Может переобучаться при неправильной настройке; Вычислительная сложность; Чувствительность к выбросам и шуму в данных	Задачи прогнозирования на табличных данных со сложными зависимостями.
LightGBM	Высокая скорость обучения по сравнению с другими бустингами; Экономичное использование памяти Эффективная работа с категориальными признаками; Поддержка параллельных и распределенных вычислений	Может переобучаться на небольших датасетах; Требует тщательной настройки гиперпараметров; Менее стабильные результаты на маленьких выборках данных.	Работа с очень большими наборами данных; Задачи с высокоразмерными признаковыми пространствами; Системы реального времени, требующие быстрого переобучения; Сценарии с преобладанием категориальных признаков.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

1.3. Разведочный анализ данных

Графики распределения для каждого признака до и после масштабирования:

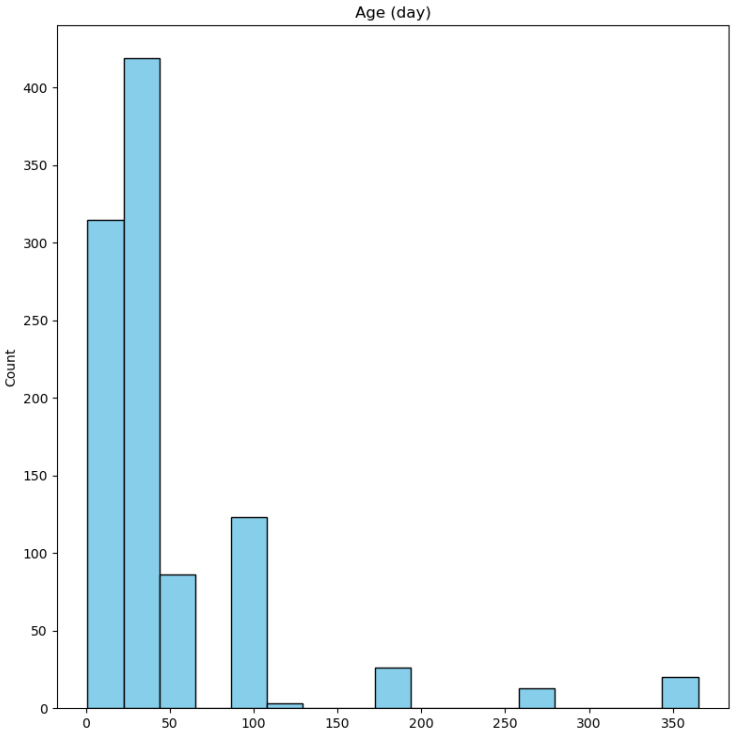


Рисунок 1 – Гистограмма распределения возраста бетона

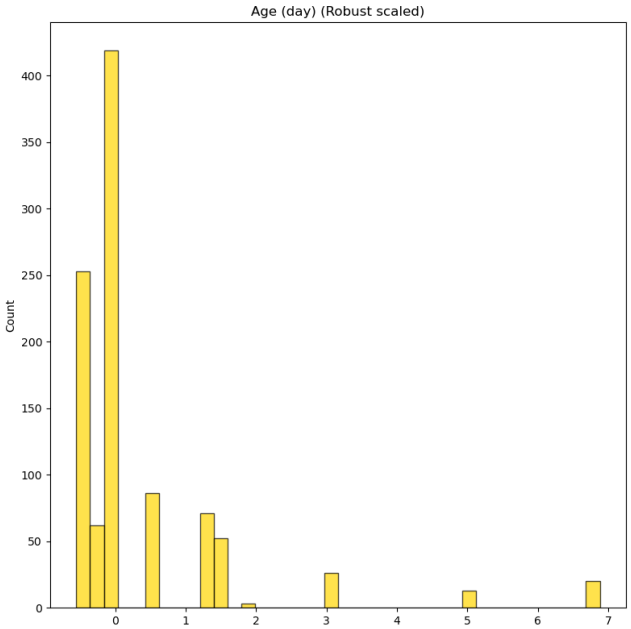


Рисунок 2 – Гистограмма распределения возраста бетона (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

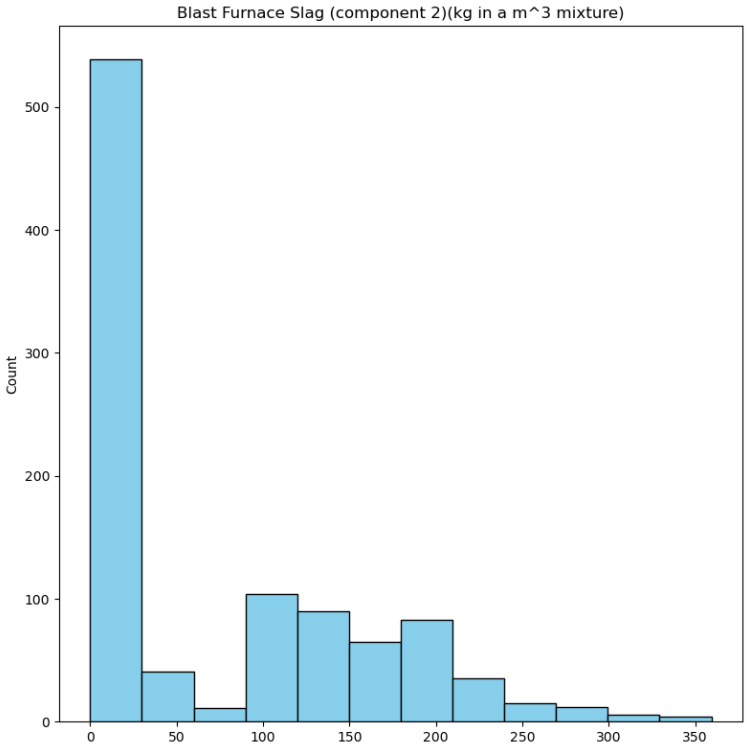


Рисунок 3 – Гистограмма распределения количества гранулированного доменного шлака

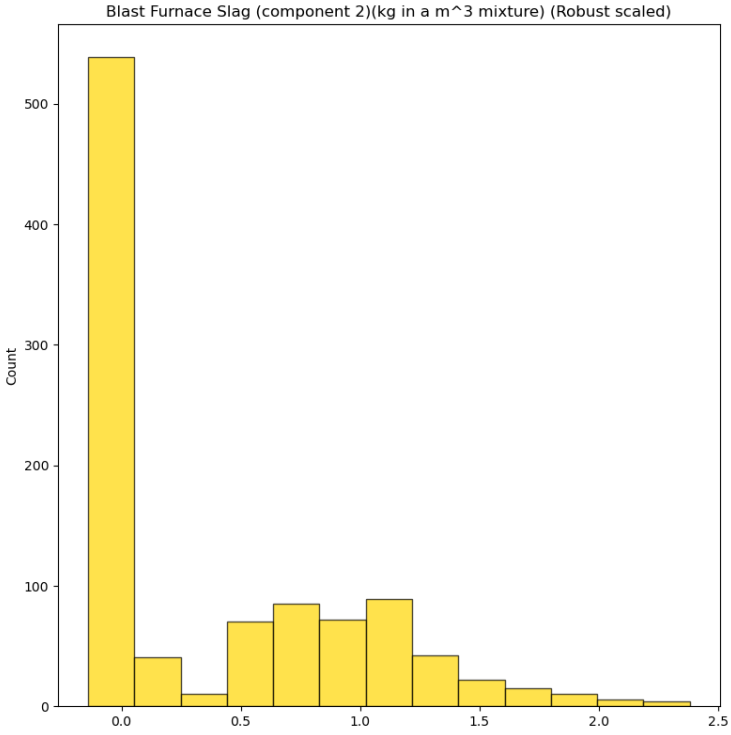


Рисунок 4 – Гистограмма распределения количества гранулированного доменного шлака (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

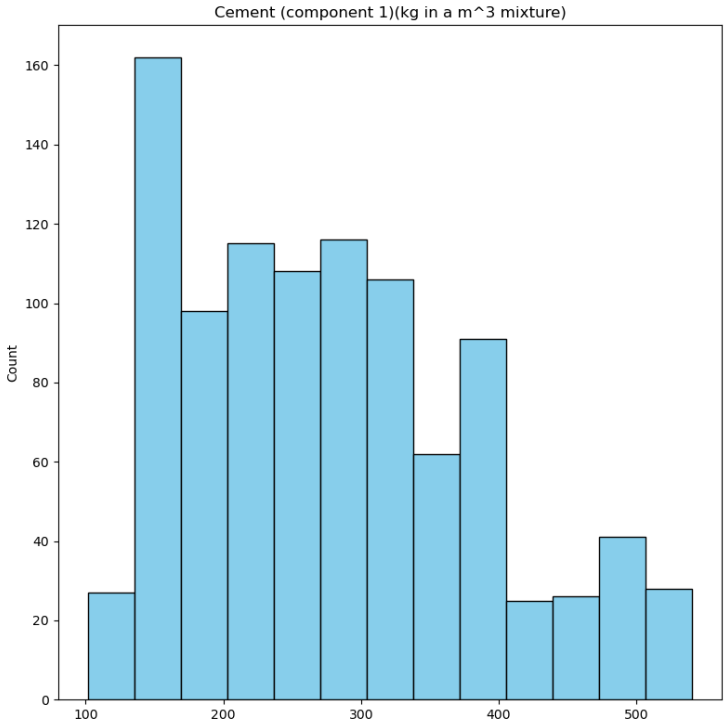


Рисунок 5 – Гистограмма распределения количества цемента

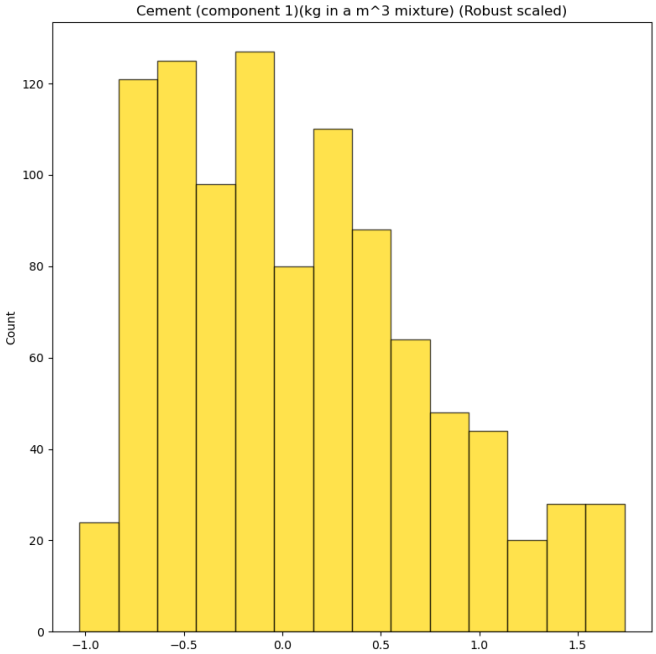


Рисунок 6 – Гистограмма распределения количества цемента (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

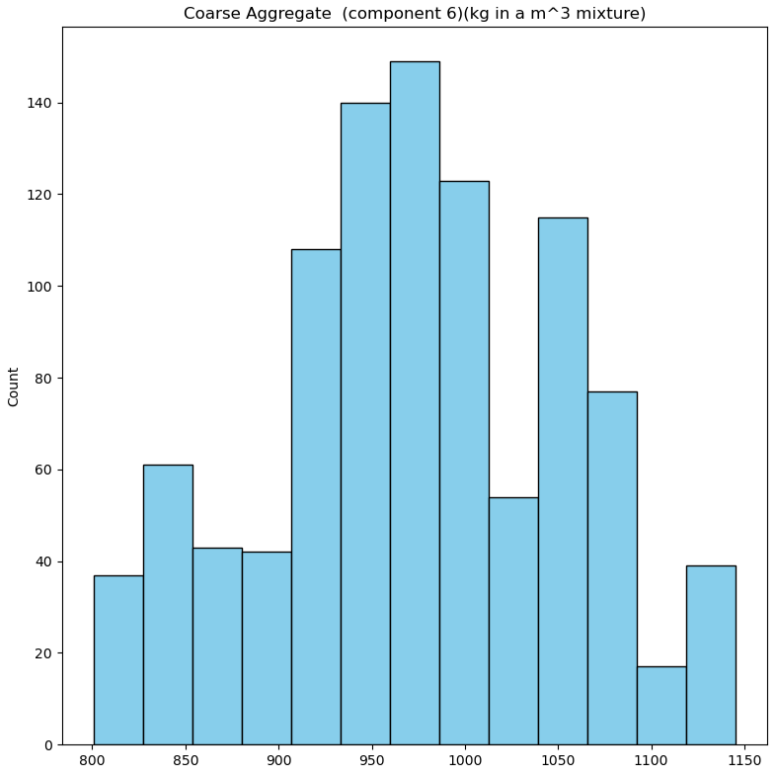


Рисунок 7 – Гистограмма распределения количества мелкого заполнителя

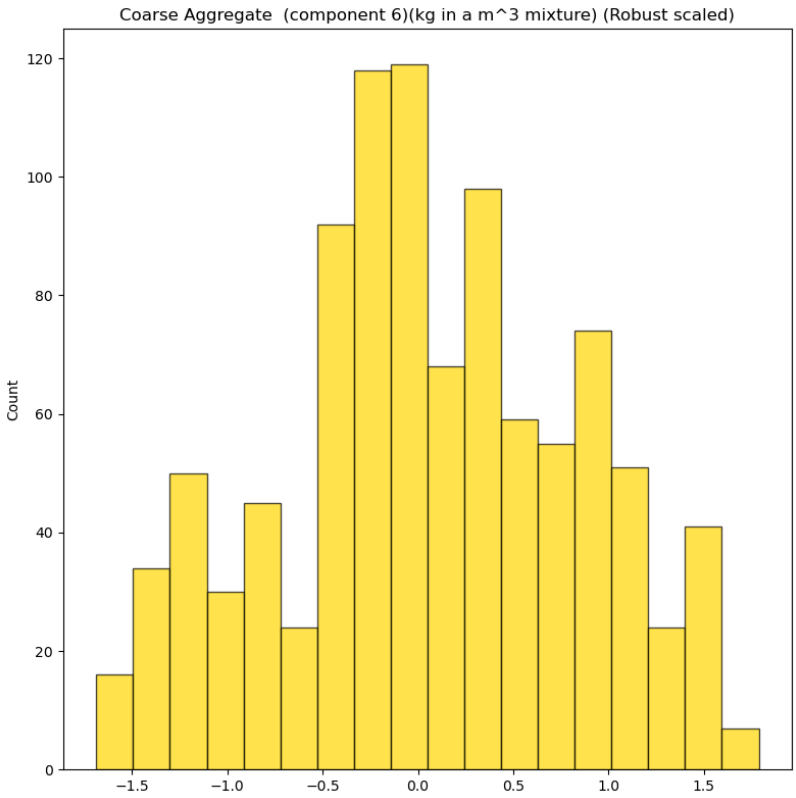


Рисунок 8 – Гистограмма распределения количества мелкого заполнителя (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

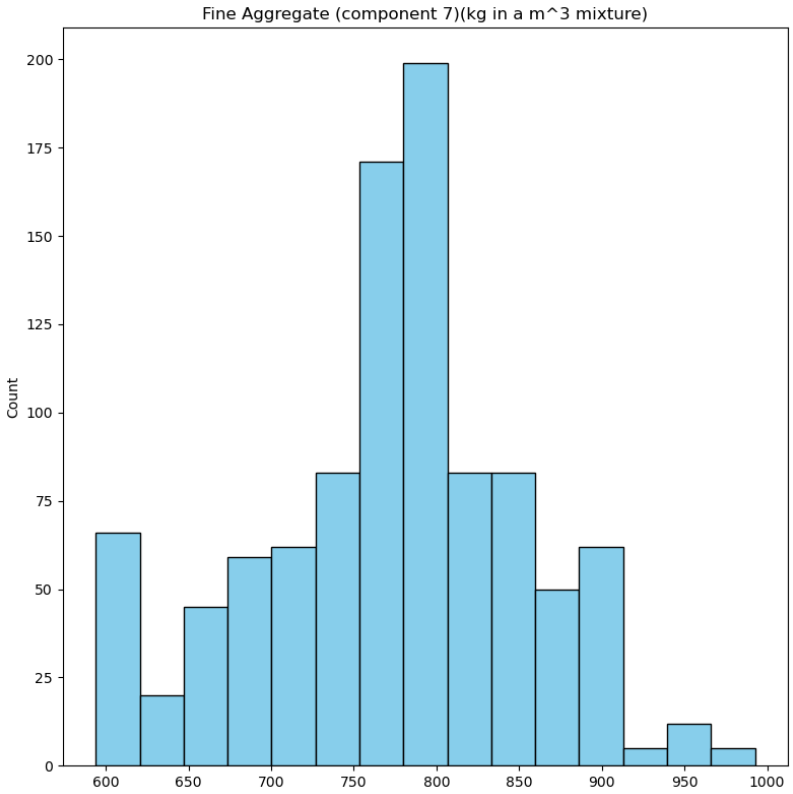


Рисунок 9 – Гистограмма распределения количества крупного заполнителя

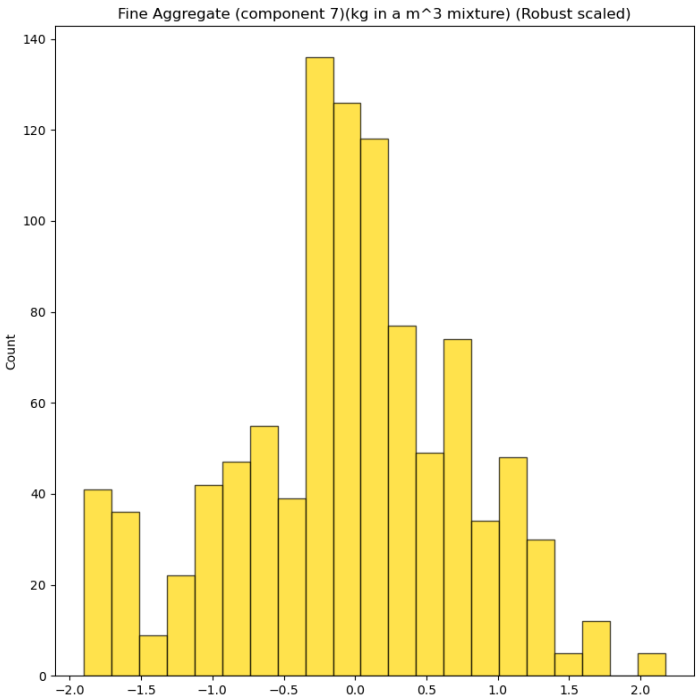


Рисунок 10 – Гистограмма распределения количества крупного заполнителя (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

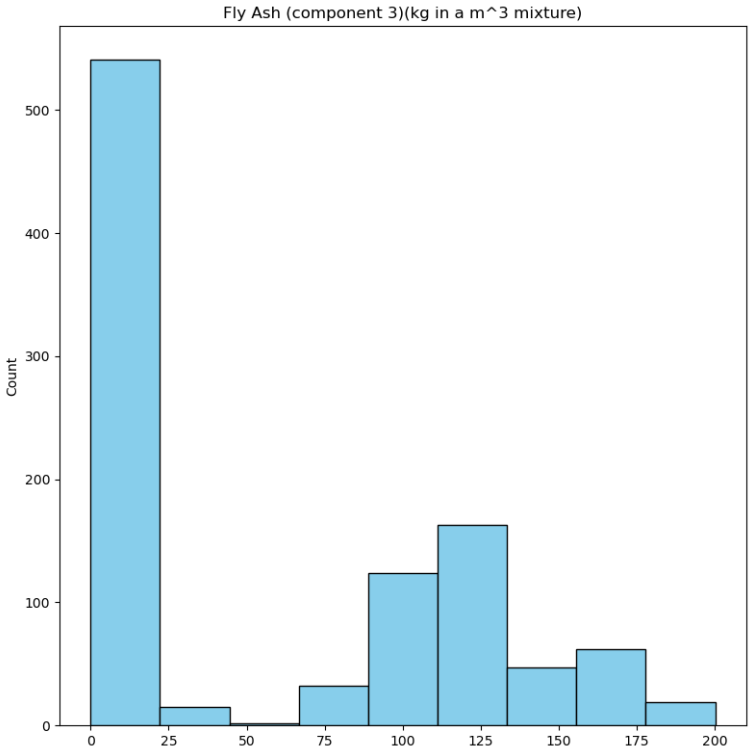


Рисунок 11 – Гистограмма распределения количества золы-уноса

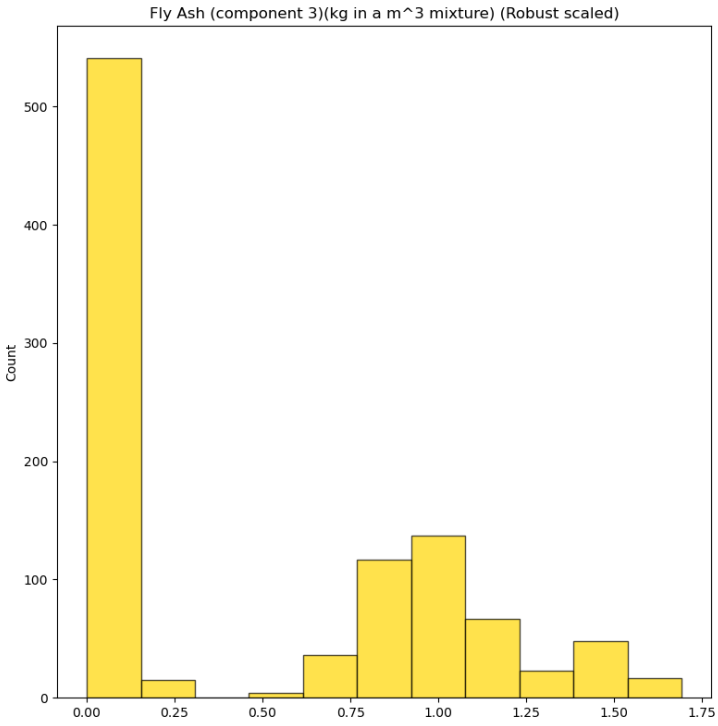


Рисунок 12 – Гистограмма распределения количества золы-уноса (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

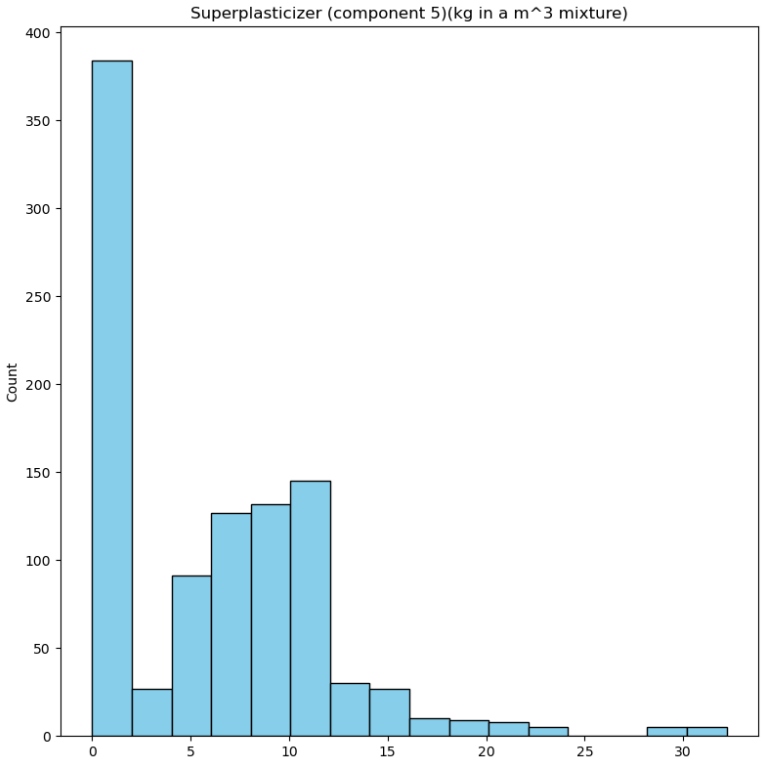


Рисунок 13 – Гистограмма распределения количества суперпластификатора

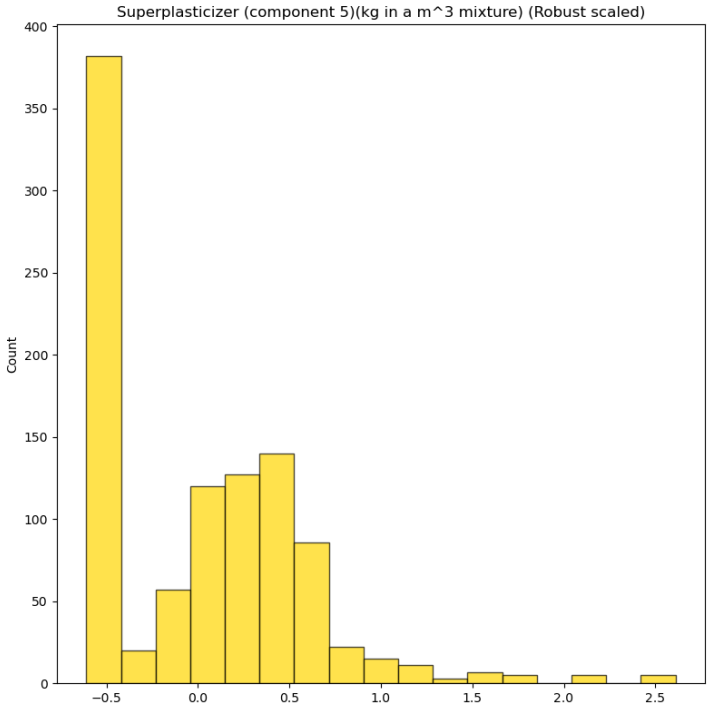


Рисунок 14 – Гистограмма распределения количества суперпластификатора (Robust Scaler)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

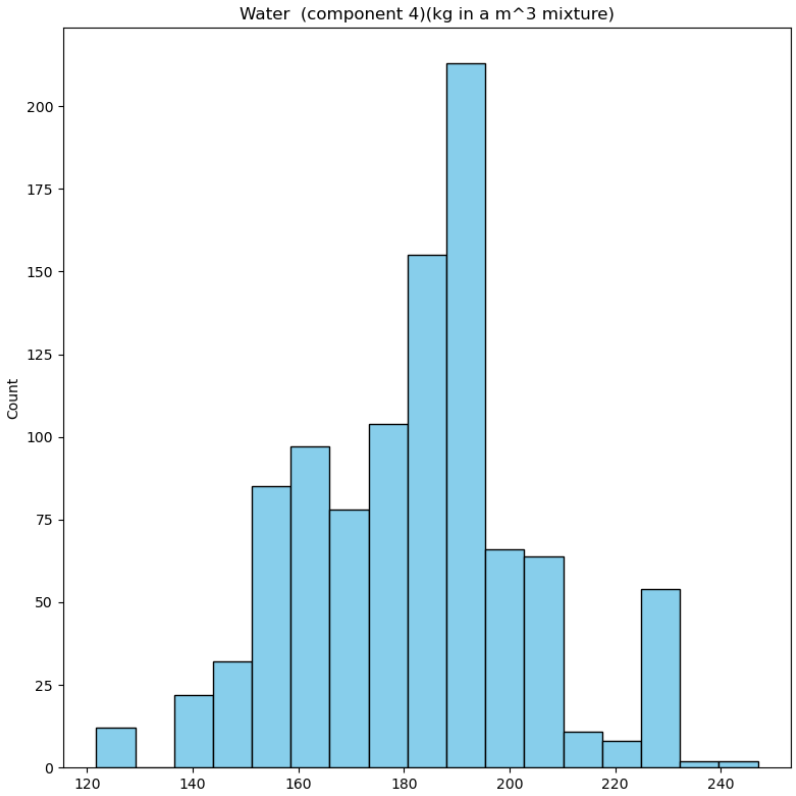


Рисунок 15 – Гистограмма распределения количества воды

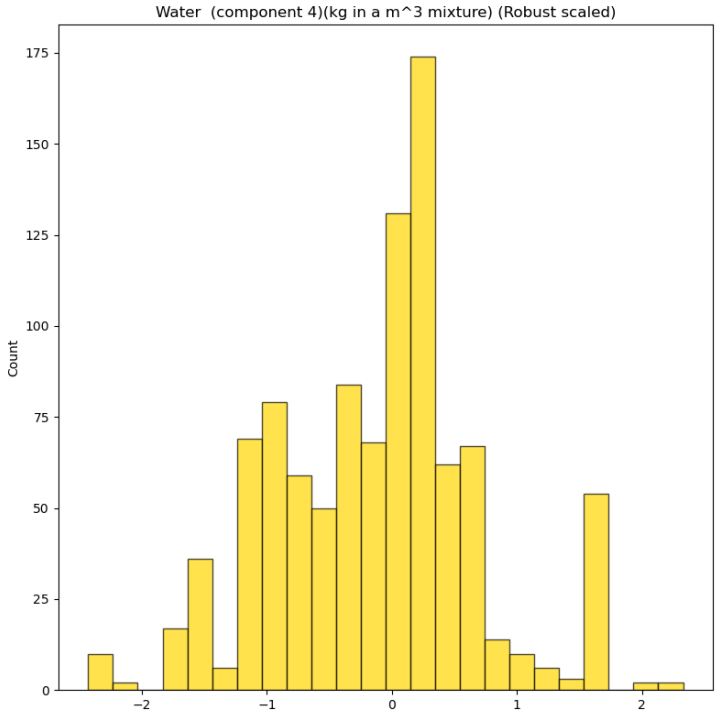


Рисунок 16 – Гистограмма распределения количества воды (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

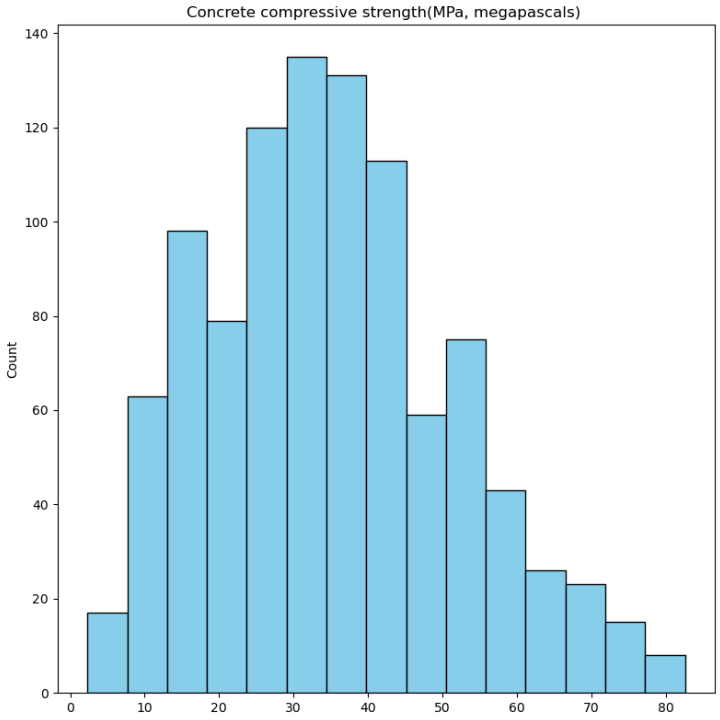


Рисунок 17 – Гистограмма распределения целевой переменной

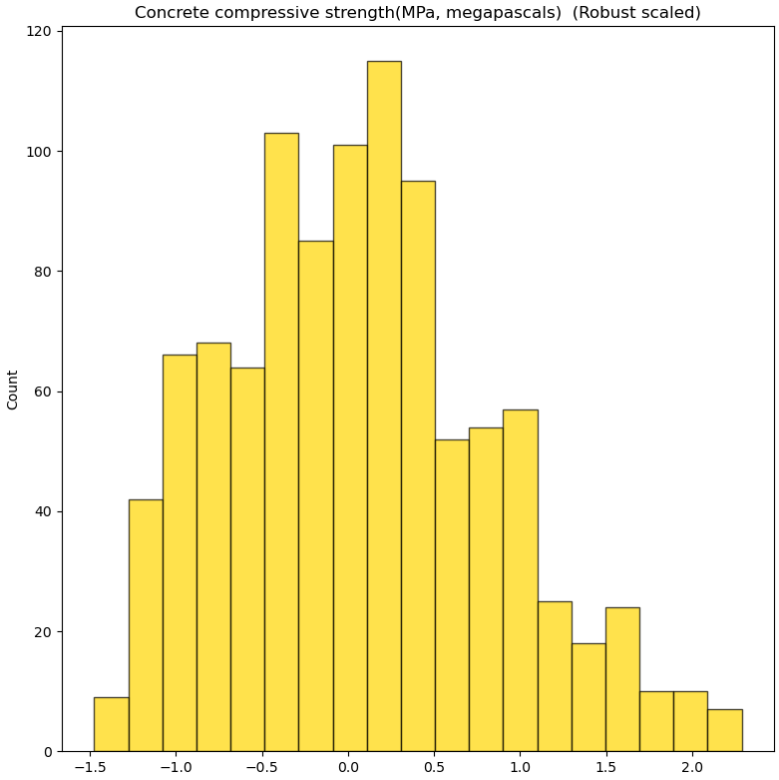


Рисунок 18 – Гистограмма распределения целевой переменной (Robust Scaled)

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Диаграммы ящик с усами:

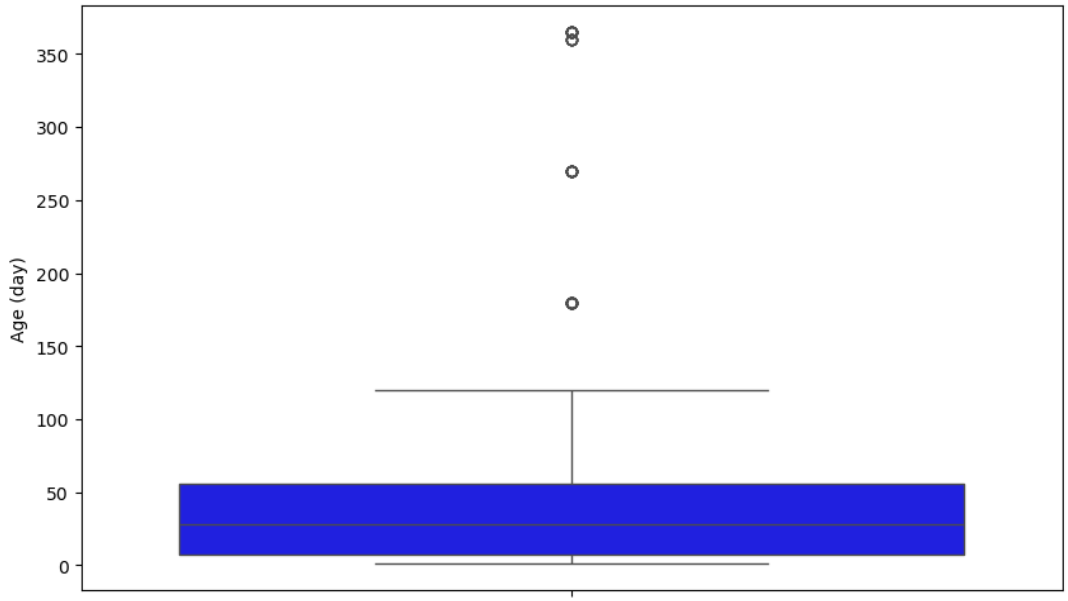


Рисунок 19 – Ящик с усами для возраста

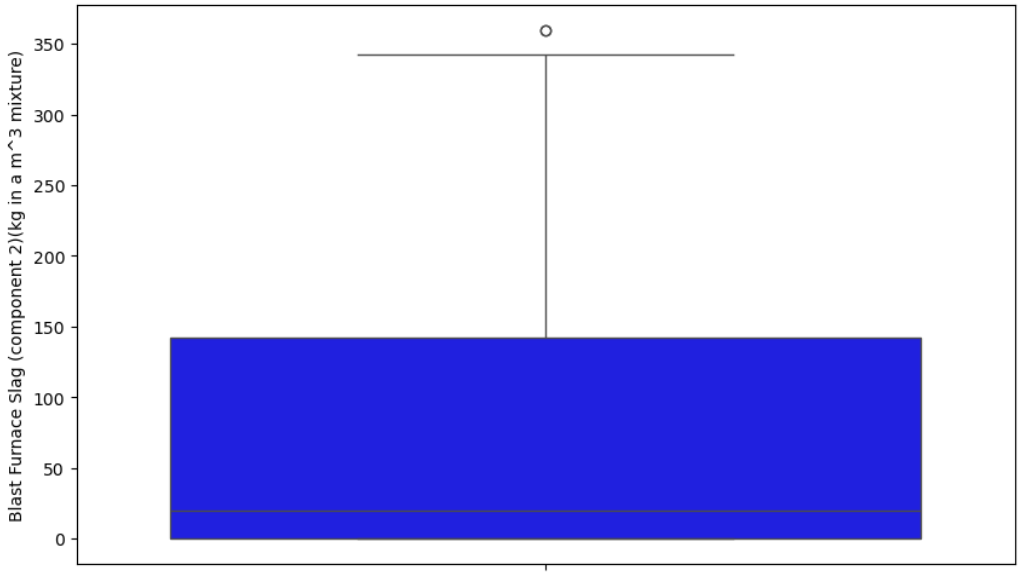


Рисунок 20 – Ящик с усами для гранулированного доменного шлака

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

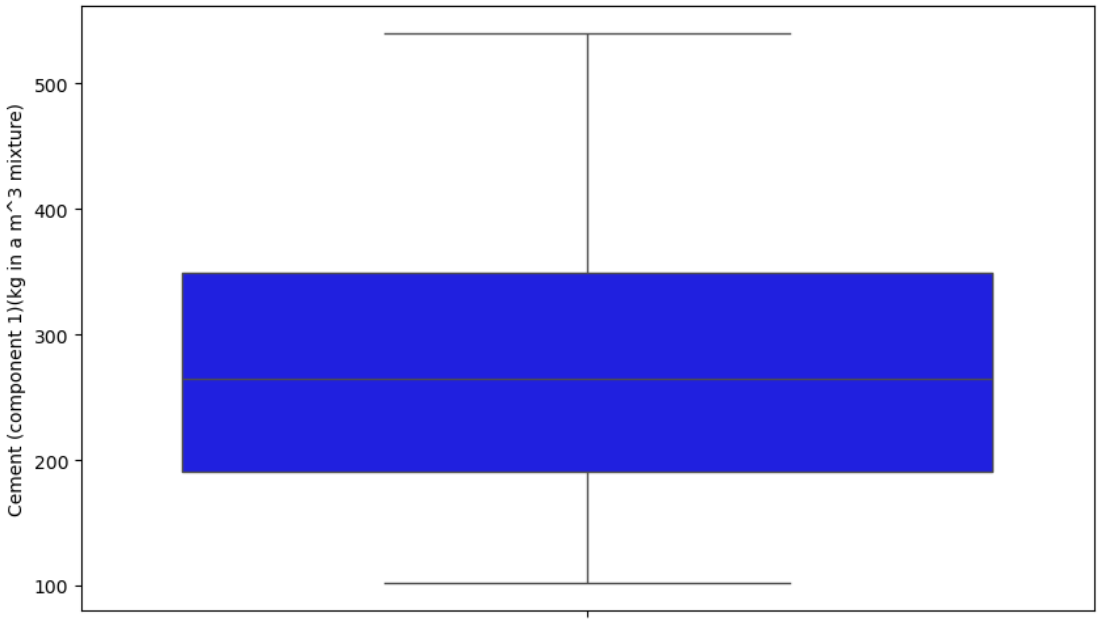


Рисунок 21 – Ящик с усами для цемента

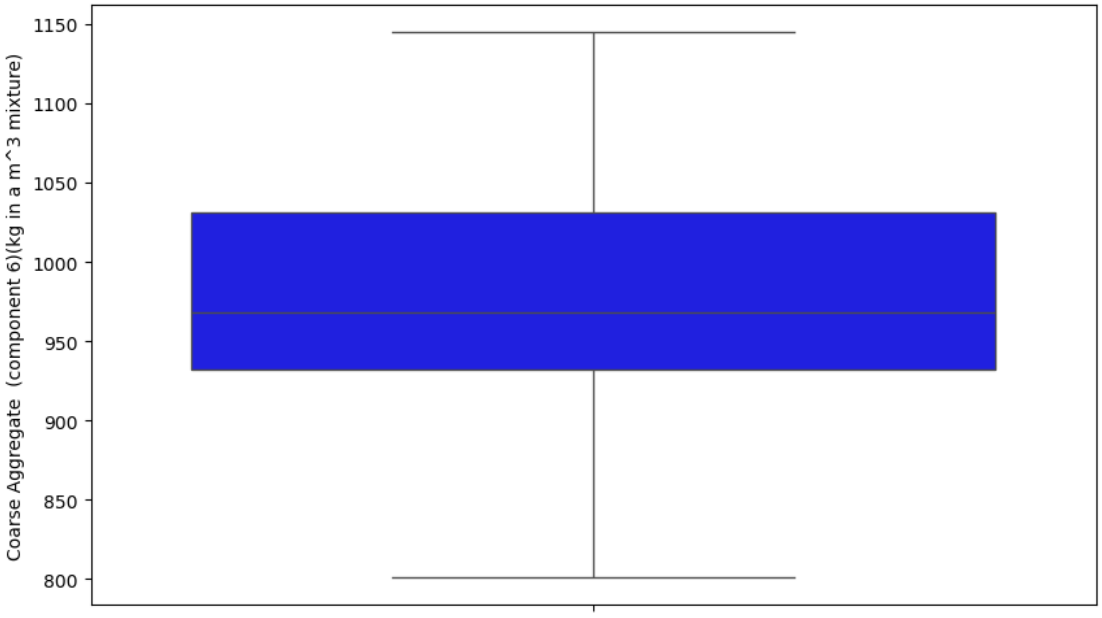


Рисунок 22 – Ящик с усами для крупного заполнителя

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

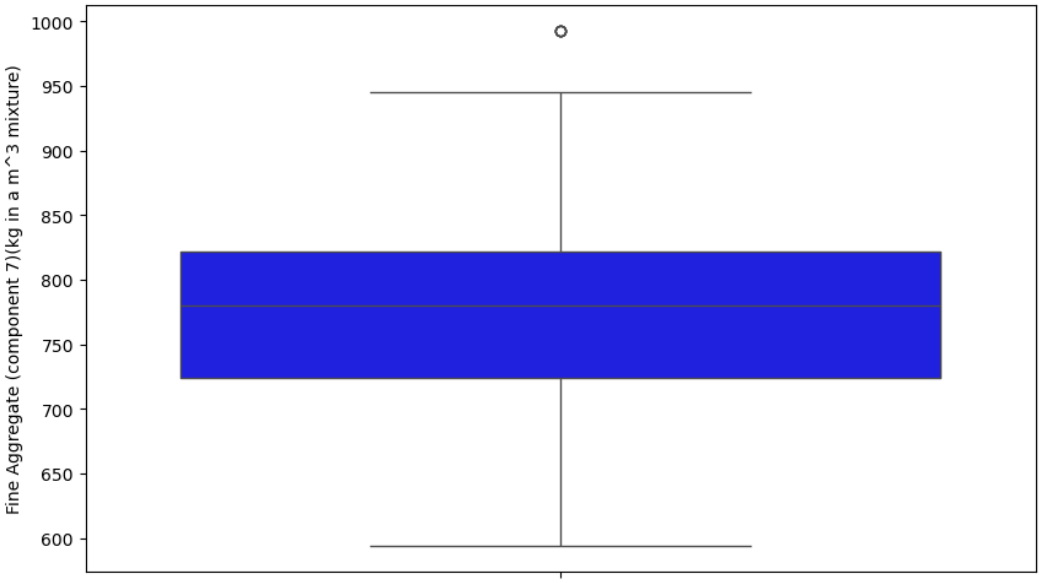


Рисунок 23 – Ящик с усами для мелкого заполнителя

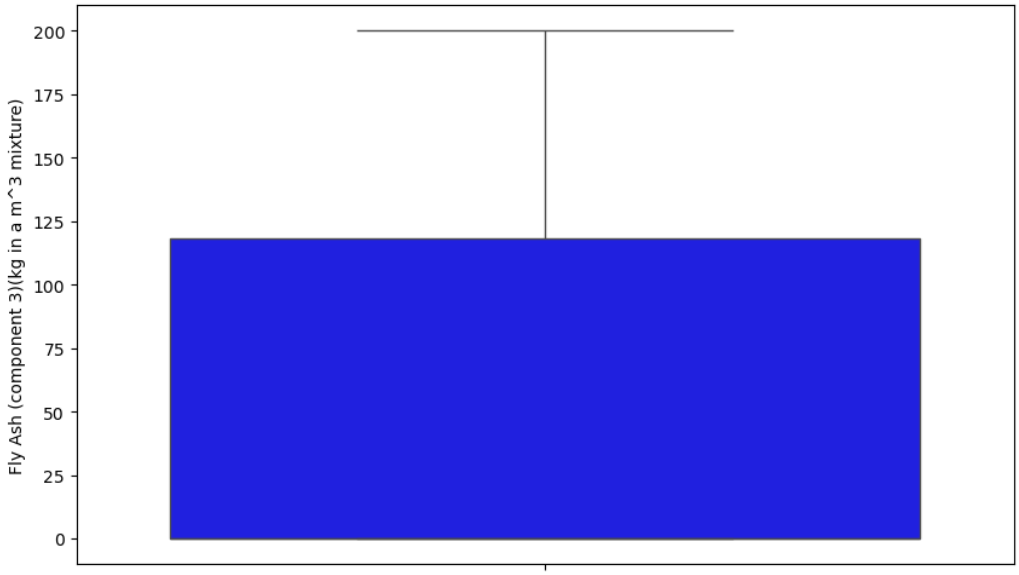


Рисунок 24 – Ящик с усами для золы-уноса

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

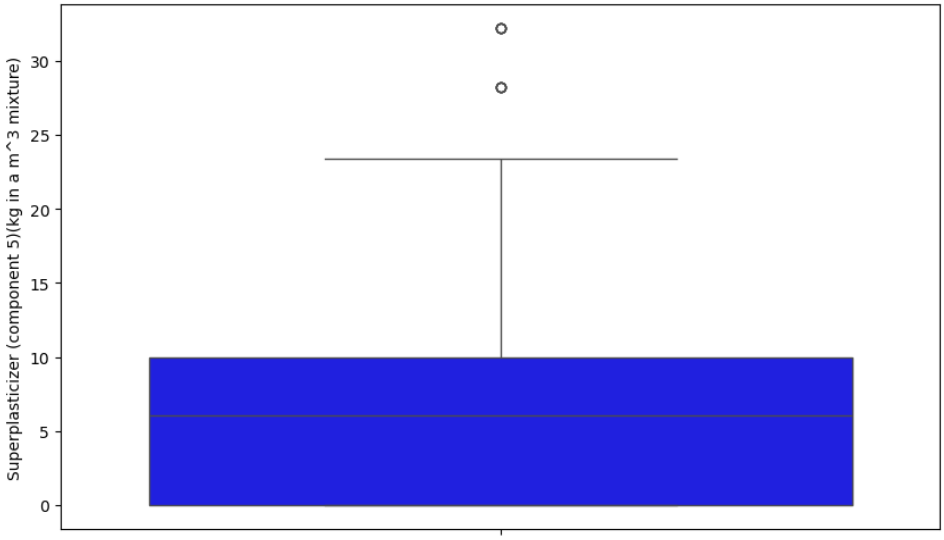


Рисунок 25 – Ящик с усами для суперпластификатора

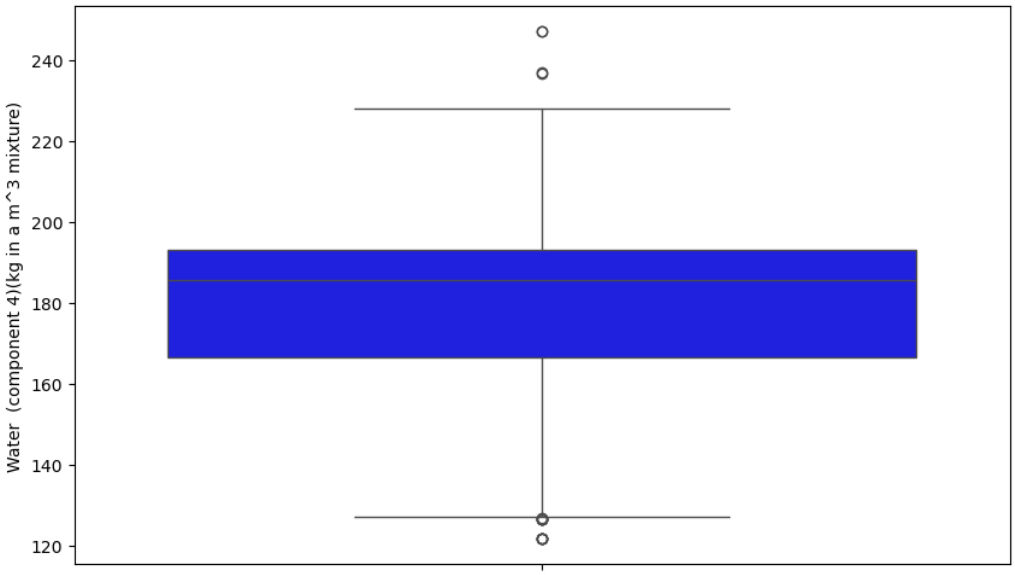


Рисунок 26 – Ящик с усами для воды

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

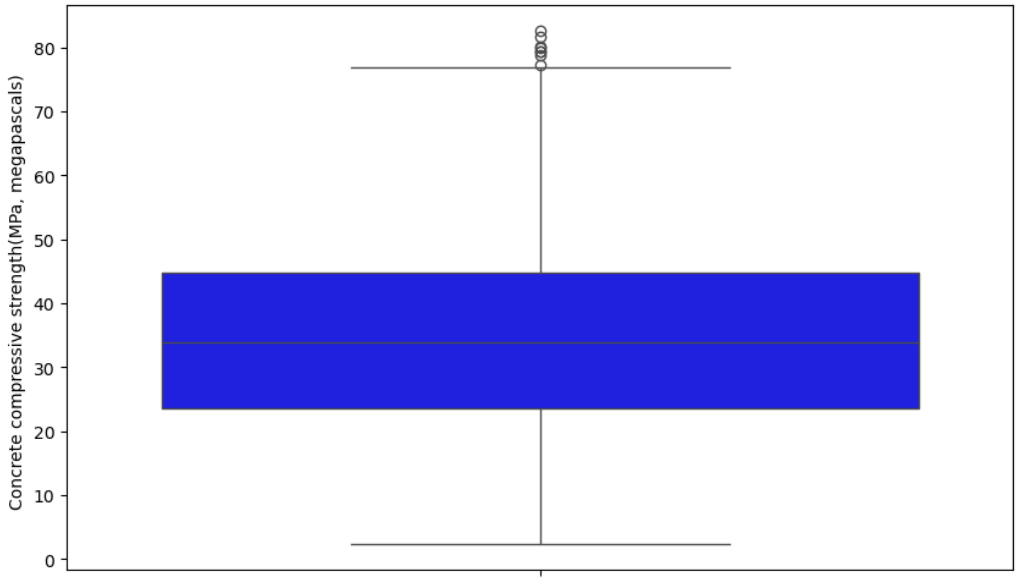


Рисунок 27 – Ящик с усами для целевой переменной

Попарные графики рассеяния:

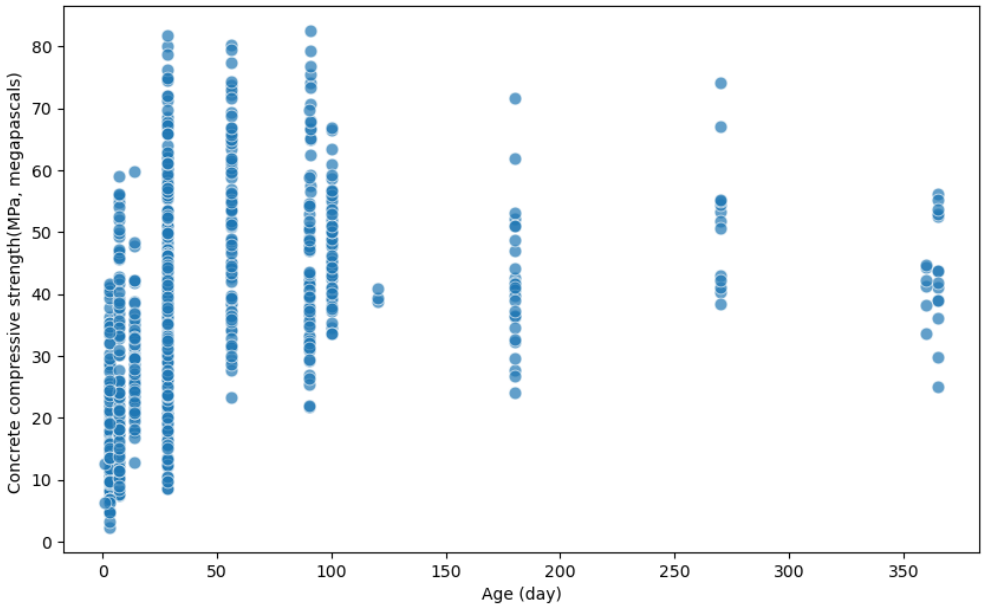


Рисунок 28 – Парный график рассеяния для возраста и целевой переменной

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

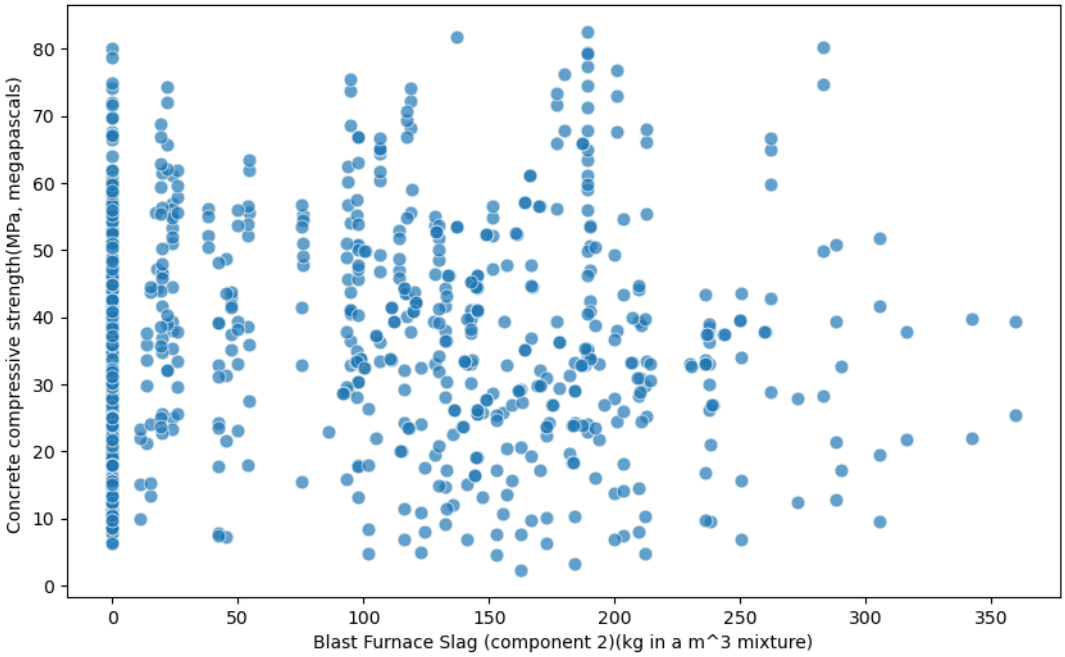


Рисунок 28 – Парный график рассеяния для шлака и целевой переменной

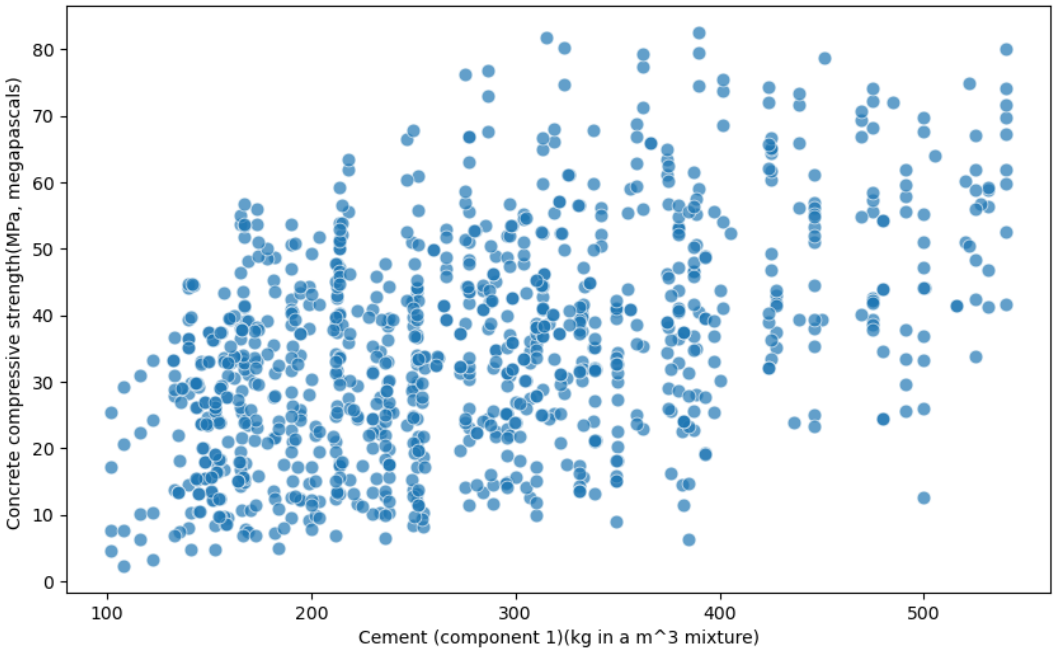


Рисунок 29 – Парный график рассеяния для цемента и целевой переменной

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

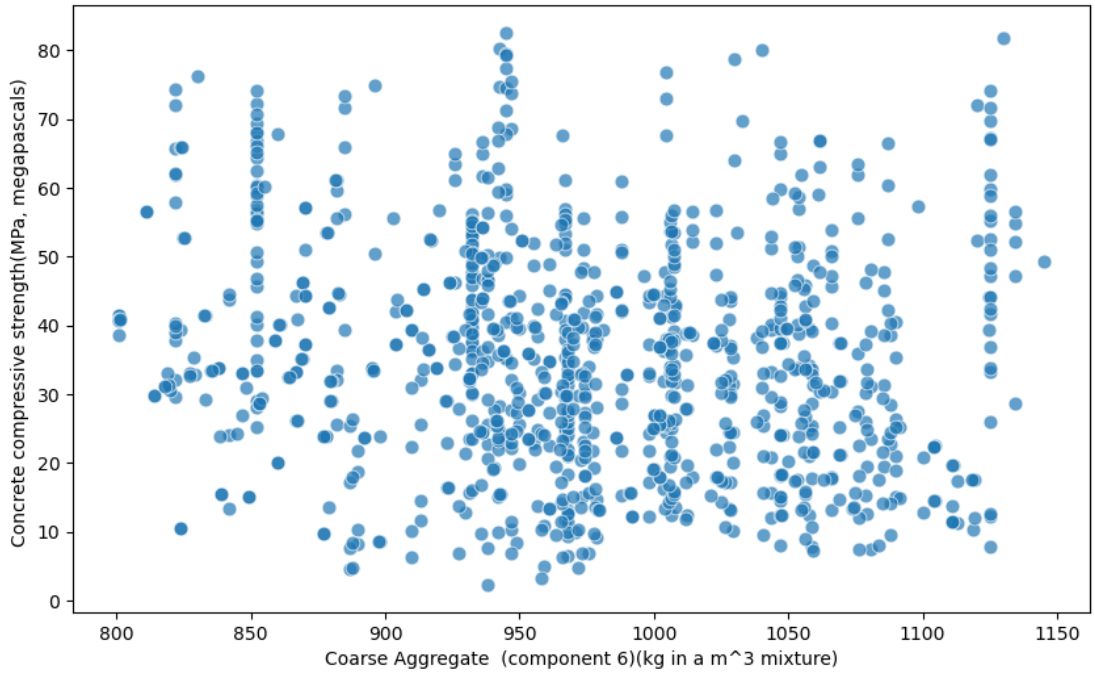


Рисунок 30 – Парный график рассеяния для крупного заполнителя и целевой переменной

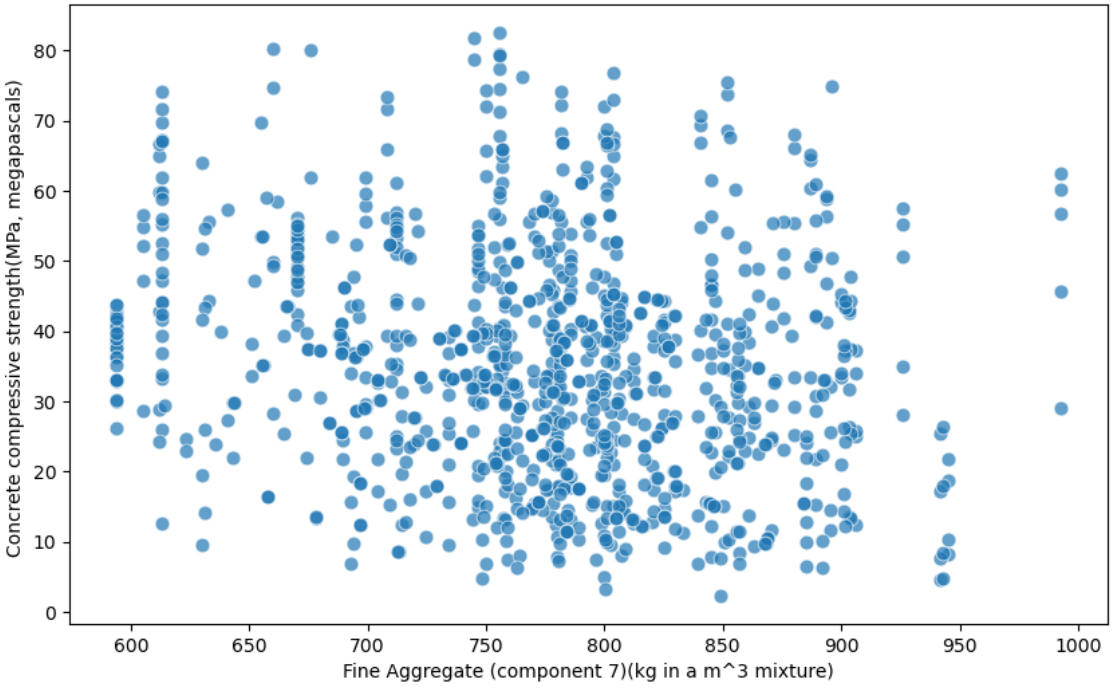


Рисунок 31 – Парный график рассеяния для мелкого заполнителя и целевой переменной

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

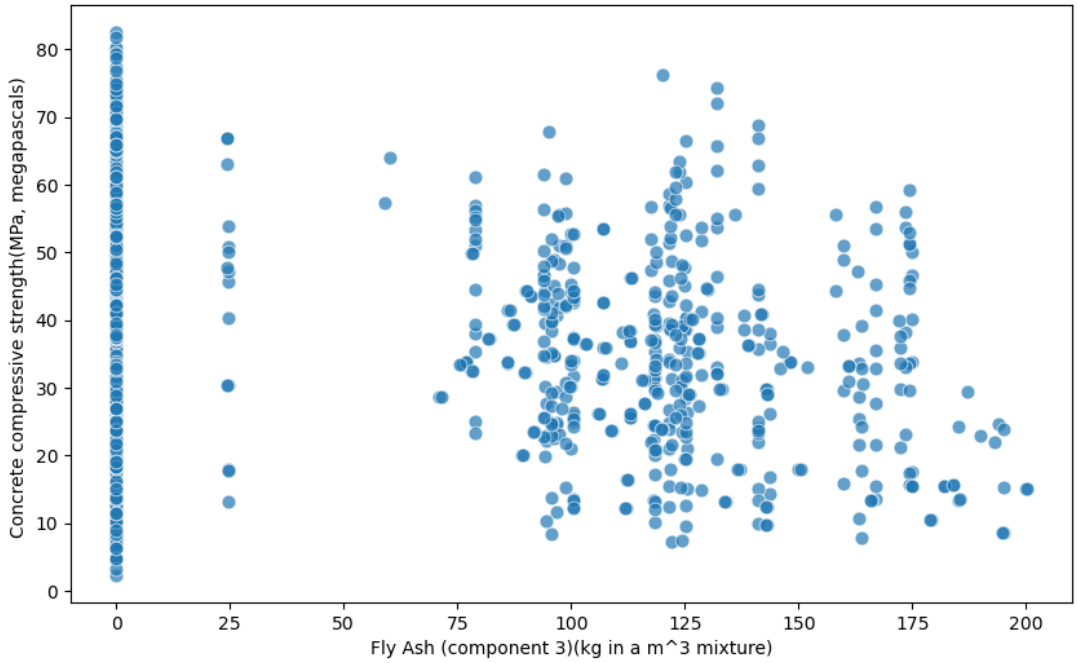


Рисунок 32 – Парный график рассеяния для мелкого золы-уноса и целевой переменной

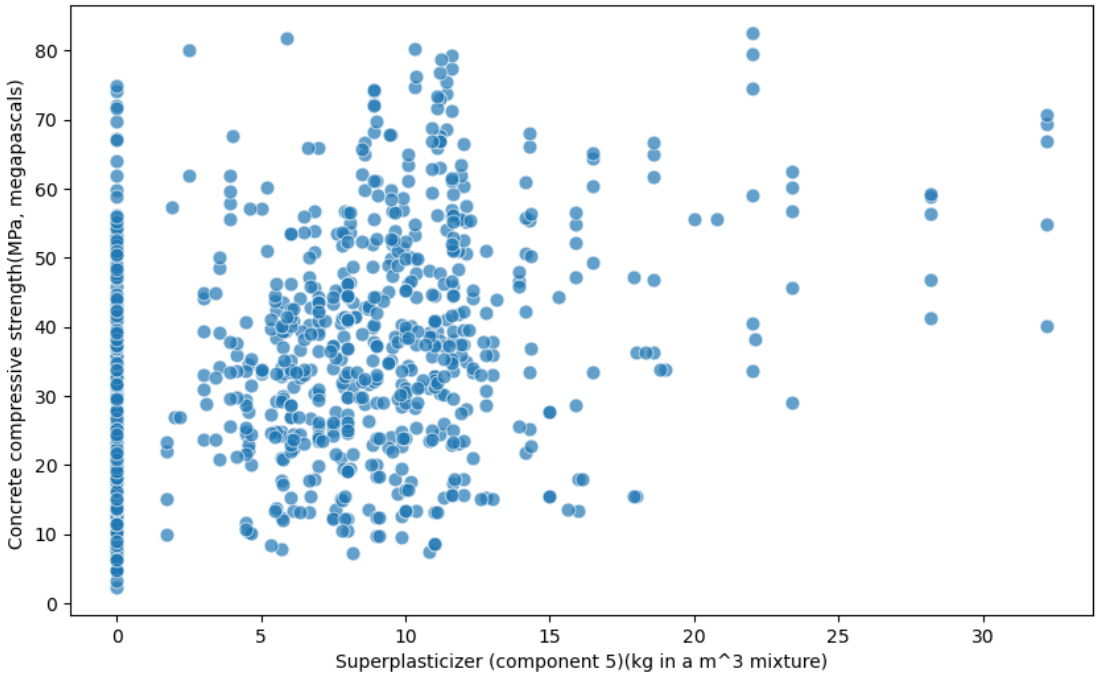


Рисунок 33 – Парный график рассеяния для суперпластификатора и целевой переменной

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

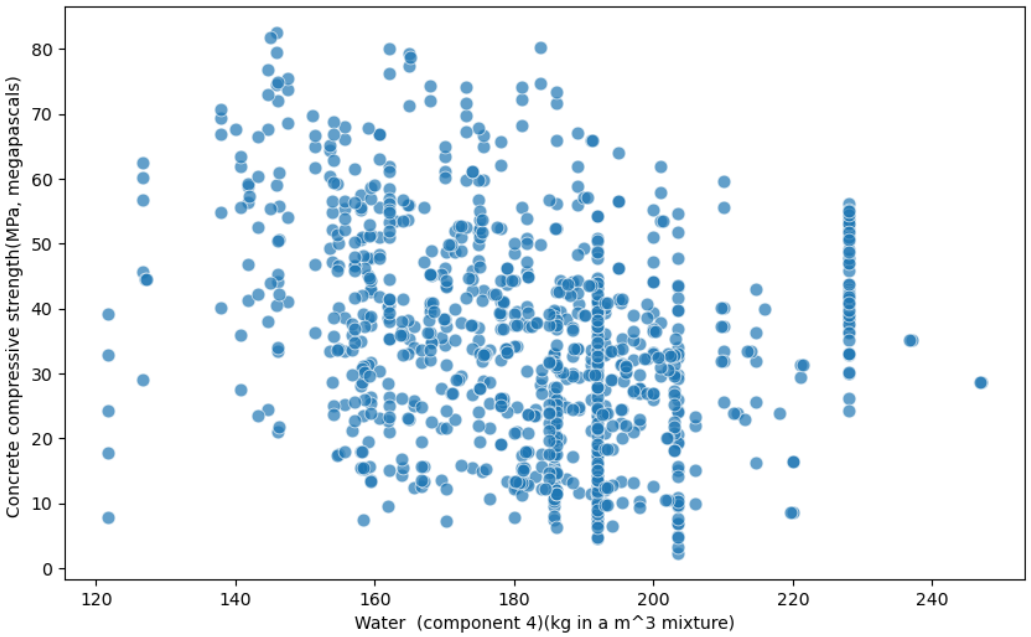


Рисунок 34 – Парный график рассеяния для мелкого воды и целевой переменной

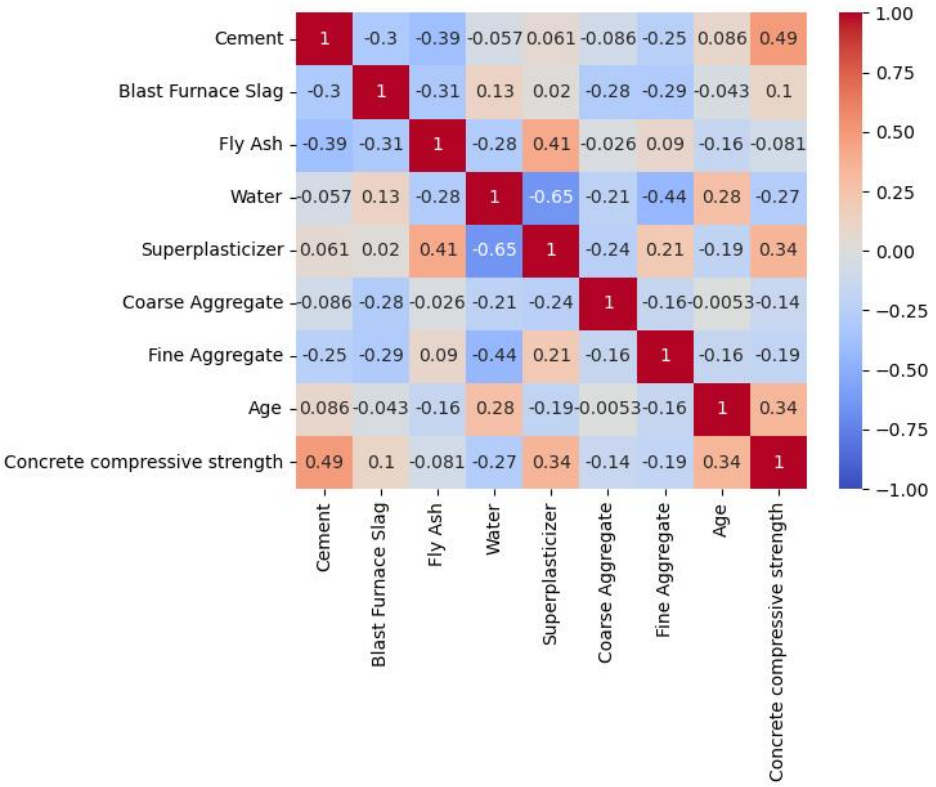


Рисунок 35 – Корреляционная матрица

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Разработка и обучение моделей

В практической части исследования были реализованы и протестированы следующие модели машинного обучения:

Линейные модели с предобработкой данных:

– Линейная регрессия (LR) - базовая линейная модель;

– Lasso-регрессия – линейная модель с L1-регуляризацией;

– ElasticNet (Enet) – комбинированная L1 и L2 регуляризация;

– Kernel Ridge Regression (KRR) - ядерный метод с регуляризацией.

Деревья и ансамбли:

– Decision Tree – дерево решений;

– Random Forest – случайный лес;

– Gradient Boosting – градиентный бустинг;

– XGBoost (XGB) – экстремальный градиентный бустинг;

– LightGBM (LGB) – легкий градиентный бустинг.

Особенности реализации

Для линейных моделей потребовалась специальная предобработка данных, реализованная с помощью make_pipeline:

Первый шаг: PowerTransformer (Yeo-Johnson) – это преобразование, которое находит оптимальный параметр λ (лямбда) для приведения распределения данных к более нормальному виду. Алгоритм автоматически подбирает такое значение λ, которое максимизирует логарифмическую функцию правдоподобия, делая распределение признаков более симметричным и близким к нормальному, что улучшает стабильность и точность линейных моделей.

Второй шаг: RobustScaler – для масштабирования данных, устойчивого к выбросам.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Эти методы обработки данных обеспечивают корректную работу линейных алгоритмов, чувствительных к масштабу и распределению данных.

2.2 Оптимизация гиперпараметров с помощью Optuna

Принцип работы Optuna:

Optuna – это фреймворк для автоматической оптимизации гиперпараметров, который использует следующие ключевые механизмы:

1. Определение пространства поиска – задаются диапазоны значений для каждого гиперпараметра.

2. Сэмплирование – интеллектуальный выбор комбинаций параметров с использованием:

– TPESampler (Tree-structured Parzen Estimator) для эффективного поиска.

– Учет истории предыдущих испытаний для направления поиска.

3. Испытания (trials) – многократное обучение модели с разными гиперпараметрами.

4. Оценка качества – каждая конфигурация оценивается по заданной метрике.

5. Прунинг – досрочное прекращение бесперспективных испытаний для экономии вычислительных ресурсов.

В данном исследовании Optuna была использована для тонкой настройки гиперпараметров таких моделей как Kernel Ridge Regression, RandomForest, XGBoost и LightGBM, что позволило значительно улучшить их прогнозную способность.

2.3 Тестирование моделей

По результатам тестирования девяти алгоритмов были получены следующие метрики качества:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

Линейные модели (LinearRegression, Lasso, ElasticNet) показали схожие результаты с R^2 около 0.78 на тестовой выборке, что указывает на их недостаточную гибкость для данного набора данных.

Деревья решений и ансамбли показали различную эффективность: DecisionTreeRegressor (R^2 = 0.84), RandomForest (R^2 = 0.90), GradientBoosting (R^2 = 0.89), XGBoost (R^2 =0.93) и LightGBM (R^2 = 0.93).

Наилучшие результаты продемонстрировали три модели: XGBoost (R^2 0.9312, RMSE 4.4478), KernelRidge (R^2 0.9287, RMSE 4.5256) и LightGBM (R^2 0.9286, RMSE 4.5305).

Для выбора финальной модели был проведен комплексный анализ, учитывающий не только метрики точности, но и специфику набора данных.

2.4 Обоснование выбора Kernel Ridge:

1. Соответствие сложности модели объему данных:

При относительно небольшом размере выборки (1000 строк) сложные ансамбливые методы типа XGBoost могут быть избыточны, тогда как KernelRidge оптимально использует доступные данные без риска излишней сложности.

2. Стабильность и надежность:

KernelRidge демонстрирует наименьший разброс результатов при кросс-валидации (±0.0160 против ±0.0193 у XGBoost), что свидетельствует о более стабильном поведении на различных подвыборках.

3. Вычислительная эффективность:

Время обучения KernelRidge составляет 0.21 секунды против 3.63 секунд у XGBoost, что делает модель более практичной для использования и возможной дальнейшей настройки.

4. Сопоставимое качество:

Разница в R^2 между KernelRidge (0.9287) и XGBoost (0.9312) составляет всего 0.0025, что статистически незначимо для практических целей, при этом KernelRidge показывает лучший показатель MAE (2.9017 против 2.9484).

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

5. Теоретическая обоснованность:

Для данных умеренной сложности ядерные методы часто оказываются более подходящими, чем сложные ансамбли, обеспечивая хороший баланс между точностью и интерпретируемостью.

Таким образом, учитывая размер набора данных, вычислительную эффективность и стабильность результатов, в качестве финальной модели выбран алгоритм KernelRidge, который демонстрирует оптимальное соотношение точности и практической применимости для решения поставленной задачи.

Таблица 4 – Сравнительная таблица метрик моделей

Модель	R^2**(CV)**	R^2**(Test)**	**MAE**	**RMSE**	**MAPE**
Linear Regression	0.7981 ± 0.0281	0.780	6.13	7.95	21.46
Lasso	0.7981 ± 0.0281	0.780	6.13	7.95	21.46
Elastic Net	0.7981 ± 0.0281	0.780	6.13	7.95	21.46
Kernel Ridge	0.9302 ± 0.0160	0.929	2.90	4.53	9.84
Decision Tree	0.7869 ± 0.0663	0.839	4.47	6.78	15.71
Random Forest	0.8914 ± 0.0239	0.905	3.51	5.23	12.77
Gradient Boosting	0.8935 ± 0.0232	0.888	4.15	5.68	14.33
XGBoost	0.9217 ± 0.0193	0.931	2.94	4.47	10.53
LightGBM	0.9228 ± 0.0219	0.929	3.06	4.53	10.90

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

2.5. Разработка приложения с графическим интерфейсом

Приложение разработано с помощью фреймворка Strtemlit, интерфейс достаточно простой: есть возможность ручного ввода для одиночного предсказания, а также есть возможность пакетного предсказания – можно загрузить файлы в формате CSV и XLSX/XLS и скачать итоговую таблицу с предсказаниями в формате CSV.

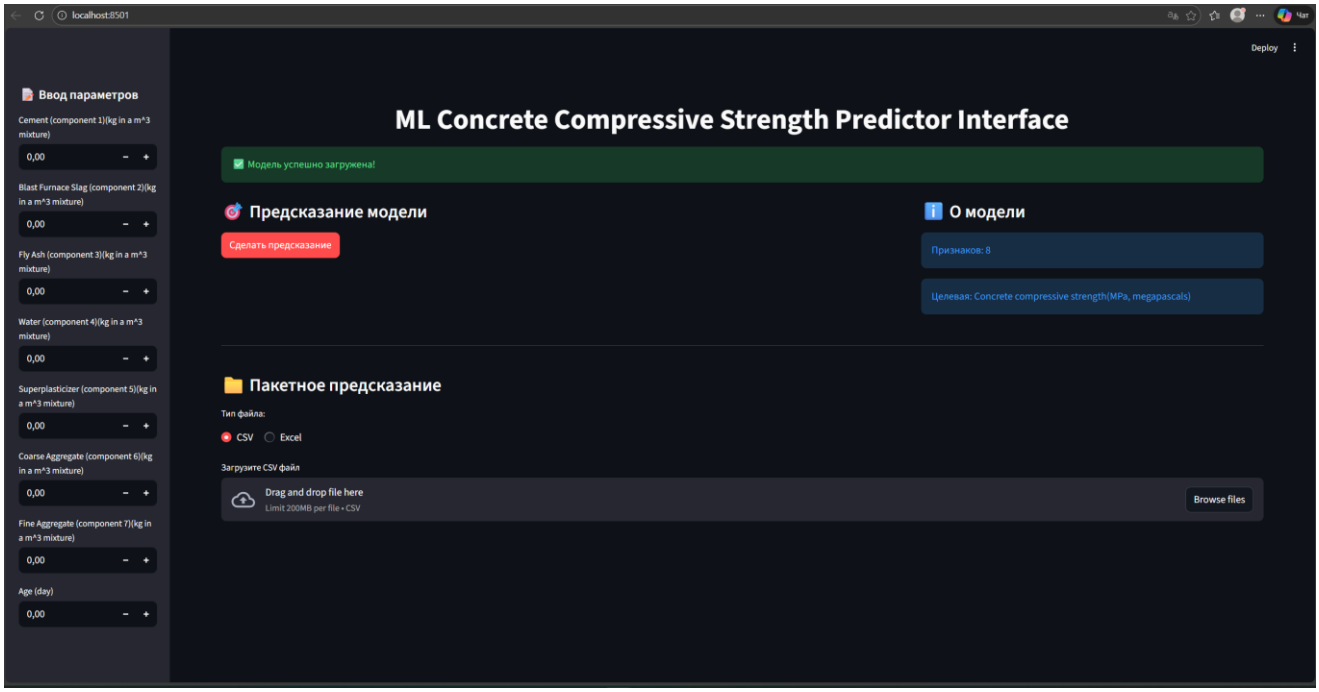
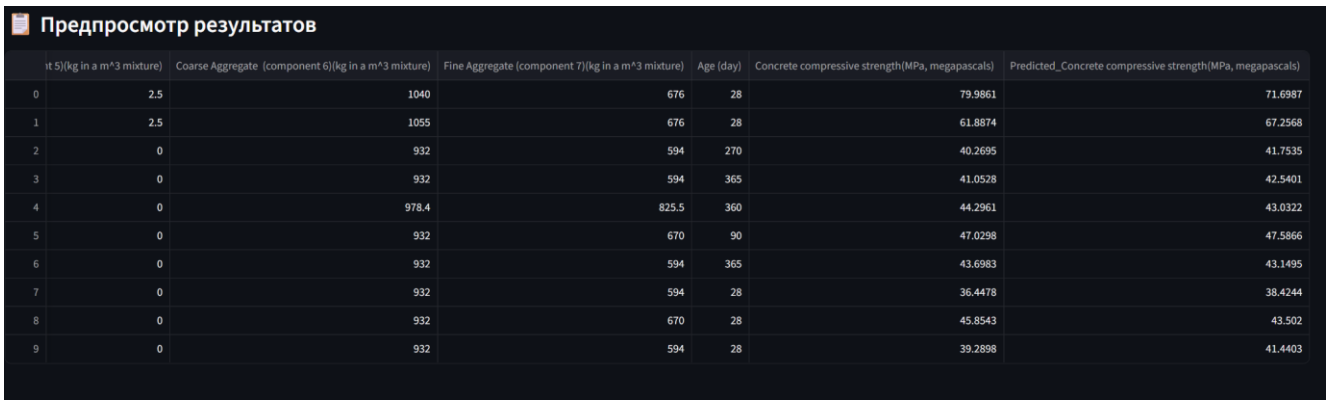


Рисунок 36 – Интерфейс веб-приложения



	it 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)	Predicted_Concrete compressive strength(MPa, megapascals)
0	2.5	1040	676	28	79.9861	71.6987
1	2.5	1055	676	28	61.8874	67.2568
2	0	932	594	270	40.2695	41.7535
3	0	932	594	365	41.0528	42.5401
4	0	978.4	825.5	360	44.2961	43.0322
5	0	932	670	90	47.0298	47.5866
6	0	932	594	365	43.6983	43.1495
7	0	932	594	28	36.4478	38.4244
8	0	932	670	28	45.8543	43.502
9	0	932	594	28	39.2898	41.4403

Рисунок 37 – Предпросмотр результатов предсказания модели

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

2.6 Создание удаленного репозитория.

Был создан удаленный репозиторий со всей кодовой базой проекта на сайте GitHub : TheSunlitMan/Concrete-Compressive-Strenght-Prediction.

Коммиты:

– Initial commit;
– Add README;
– Add data;
– Fixes.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

ЗАКЛЮЧЕНИЕ

После создания и тестирования десяти моделей машинного обучения, были получены следующие результаты на небольшом структурированном наборе данных Kernel Ridge Regression может превосходить градиентный бустинг на основе решающих деревьев как по точности предсказания, так и по скорости обучения (0.25 секунд против 3.5 секунд).

Лучшая модель была успешно интегрирована в веб-приложение с графическим интерфейсом.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 187-85-85, do@bmstu.ru
do.bmstu.ru

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Жерон, А. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем / А. Жерон; [пер. с англ.]. – СПб.: ООО «Альфа-книга», 2018. – 688 с.: ил.

2. Грас, Дж. Data Science. Наука о данных с нуля / Дж. Грас; [пер. с англ.]. - 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021. – 416 с.: ил.

3. Рашка, С. Машинное обучение с использованием Scikit-Learn и TensorFlow / С. Рашка; [пер. с англ.]. – М.: ДМК Пресс, 2018. – 452 с.: ил.