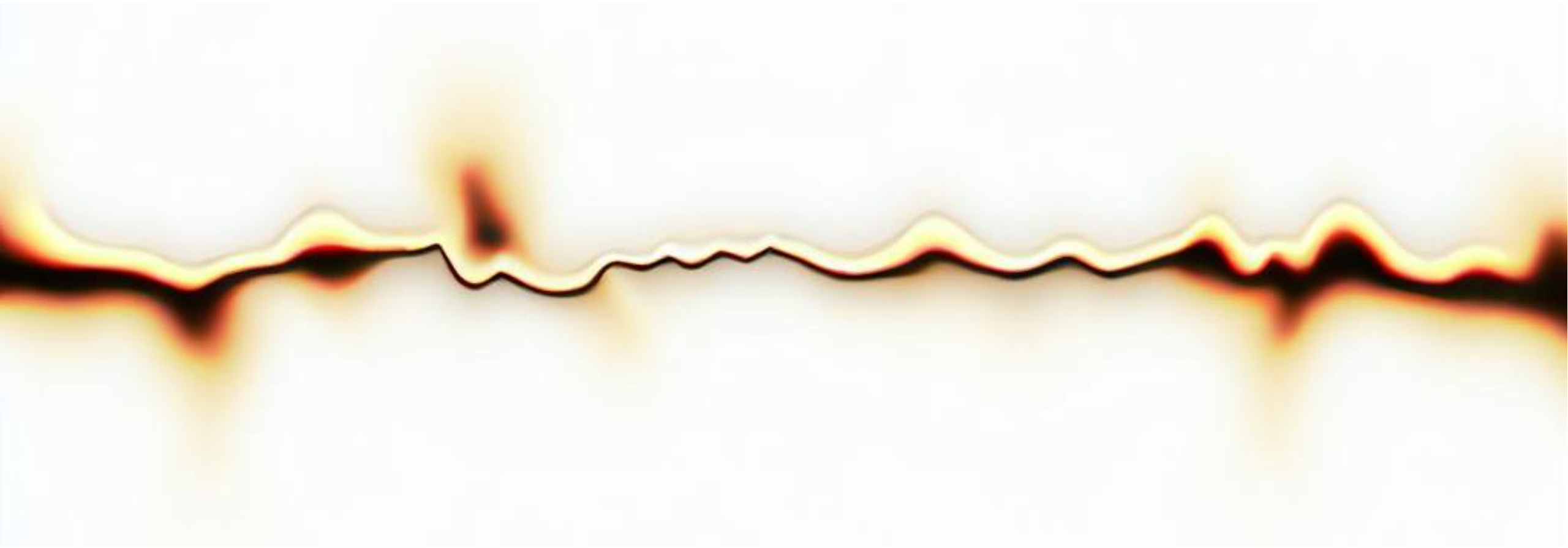



# оптимизация производственных расходов за счёт контроля температуры стали

---

Солин Михаил







## металлургический комбинат для оптимизации производственных расходов планирует уменьшить энергопотребление на этапе обработки стали

Дополнительный контроль температуры сплава стали до её передачи на доводку позволит комбинату снизить затраты на энергопотребление и увеличит срок эксплуатации оборудования за счёт сокращения износа оборудования

Для реализации поставленной цели нам необходимо написать модель, которая предскажет конечную температуру стали с ошибкой МАЕ не больше 6 градусов.

Ковш, не подходящий под выделенную температуру, будет отбракован и не отправится на доводку.

Модель должна быть интерпретируемой, т.к. комбинат должен понимать влияние состава материала на температуру стали, а значит нейросети для формирования модели не подходят, только интерпретируемые алгоритмы.



# имеющиеся в нашем распоряжении данные:

- **arc.csv** — данные об электродах
- **bulk.csv** — данные о подаче сыпучих материалов (объём)
- **bulk\_time.csv** — данные о времени подачи сыпучих материалов
- **gas.csv** — данные о продувке сплава газом
- **temp.csv** — результаты измерения температуры
- **wire.csv** — данные о проволочных материалах (объём)
- **wire\_time.csv** — данные о времени подачи проволочных материалов



# контрольные метрики:

## ОСНОВНАЯ: MAE (Mean Absolute Error)

**должна быть не ниже 6°C**

средняя абсолютная ошибка измеряет среднее абсолютное отклонение между предсказанными и фактическими значениями. Она покажет нам ошибку (отклонение) в градусах по Цельсию

## ДОПОЛНИТЕЛЬНАЯ: RMSE (Root Mean Squared Error)

**должна быть близка к MAE**

корень из среднеквадратической ошибки более чувствителен к выбросам и аномалиям в данных

использование обеих метрик позволяет получить более полное представление о качестве модели и ее способности обобщать данные

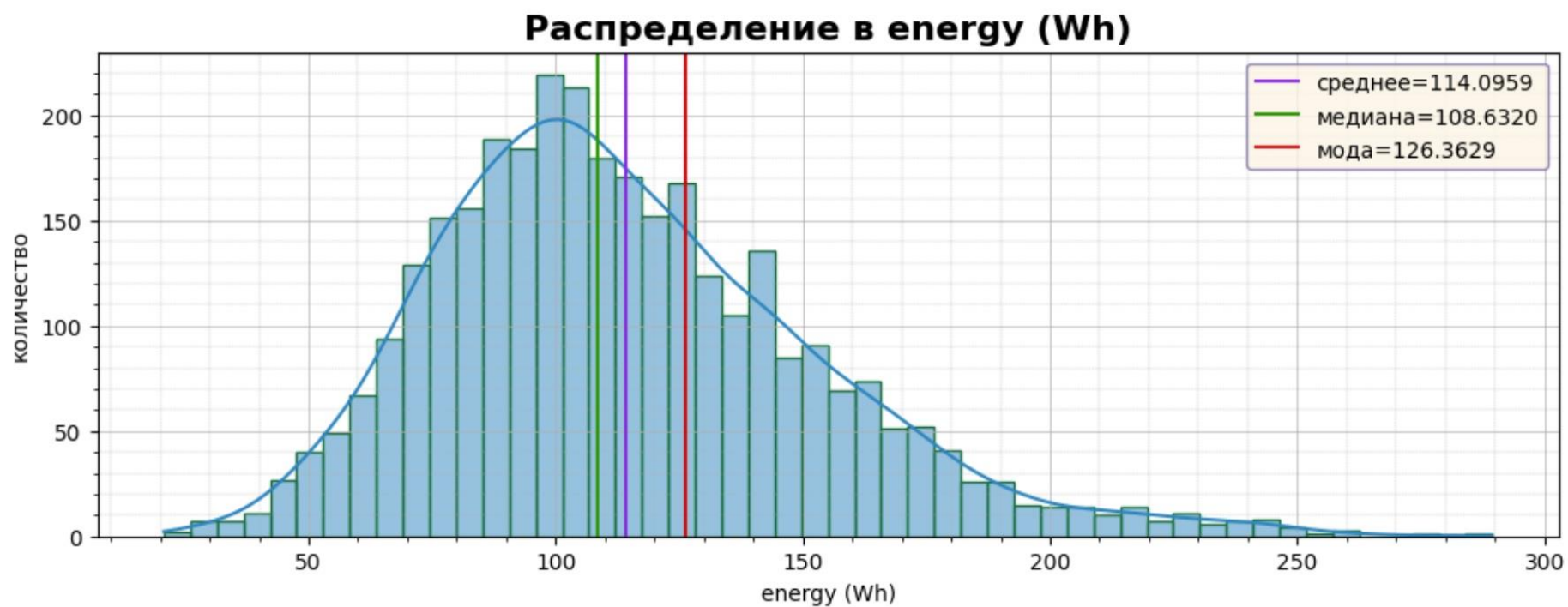


# пошаговый план для реализации цели:

- исследовательский анализ данных
- подготовка данных к машинному обучению
- перебор нескольких алгоритмов машинного обучения и их ансамблей
- выбор лучшей модели
- проверка выбранной модели на адекватность
- сохранение модели



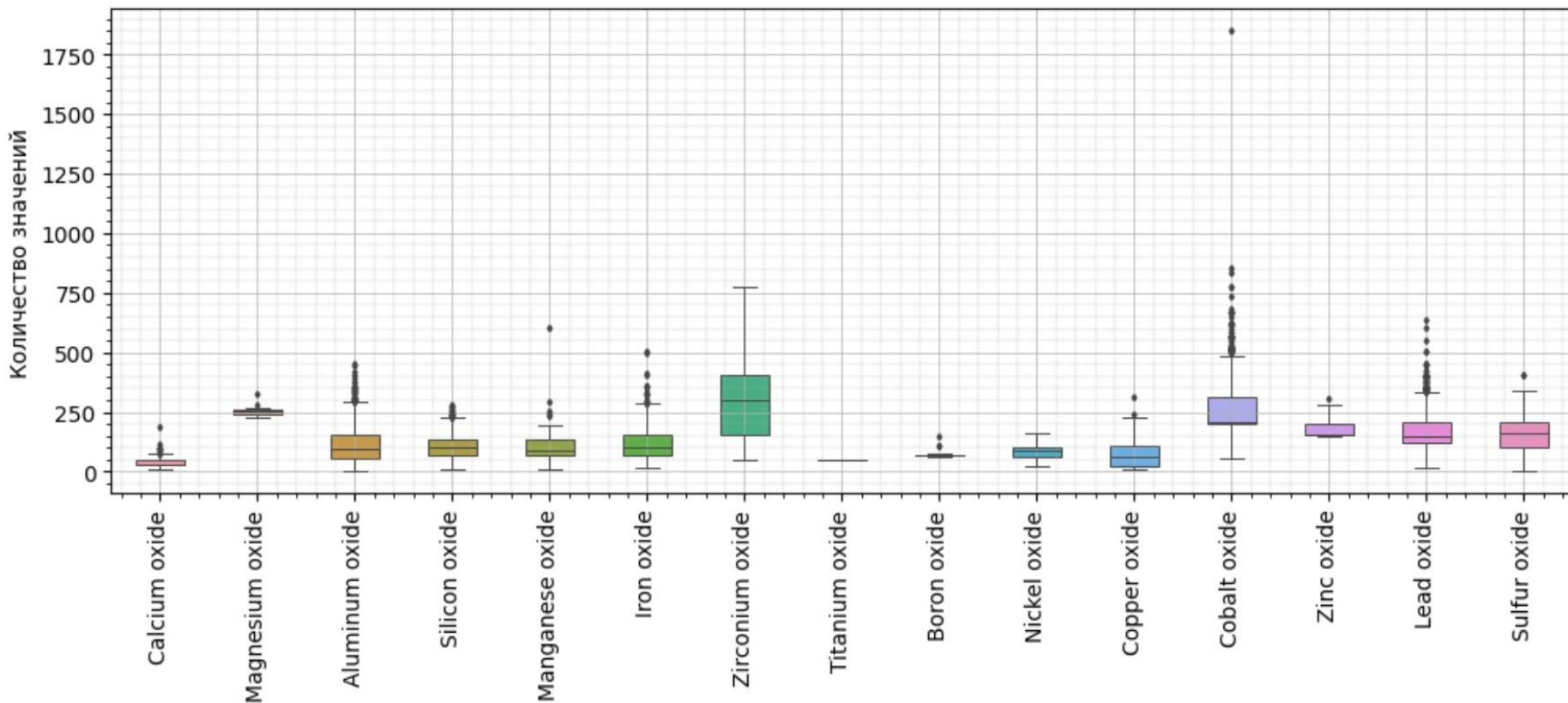
# данные об электродах



- исправили формат дат
- удалили явный выброс в *Reactive power*
- рассчитали энергию в ватт-часах (Wh), которая тратится на плавку, удалили в ней затраты свыше 300 Wh, посчитав их за выбросы



# данные о сыпучих материалах (объём)



- изучили датасет
- отметили как потенциальные к удалению те признаки, в которых количество пропусков превышает 90%
- сами пропуски заполнили нулём

# данные о сыпучих материалах (время)

Количество значений в каждом признаке для df\_bulk\_time

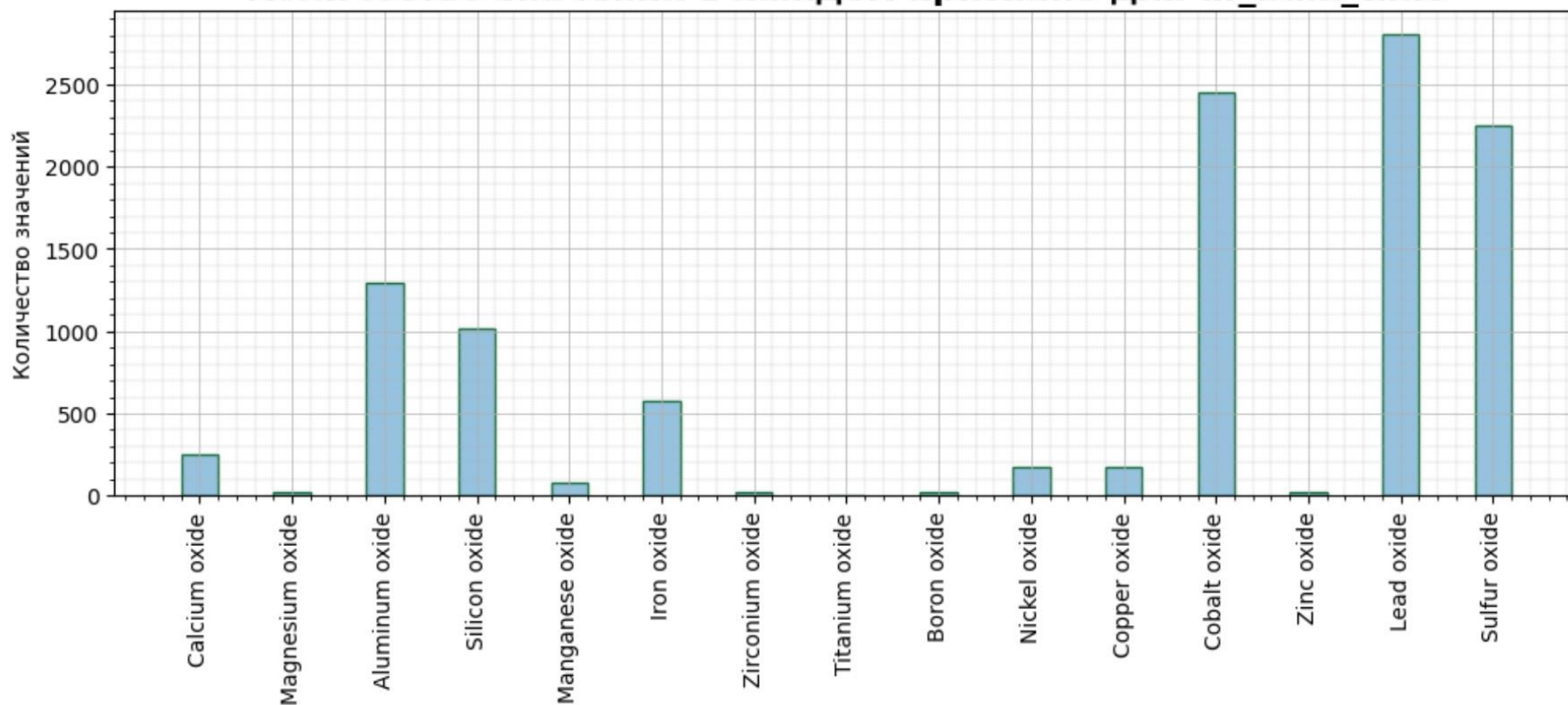
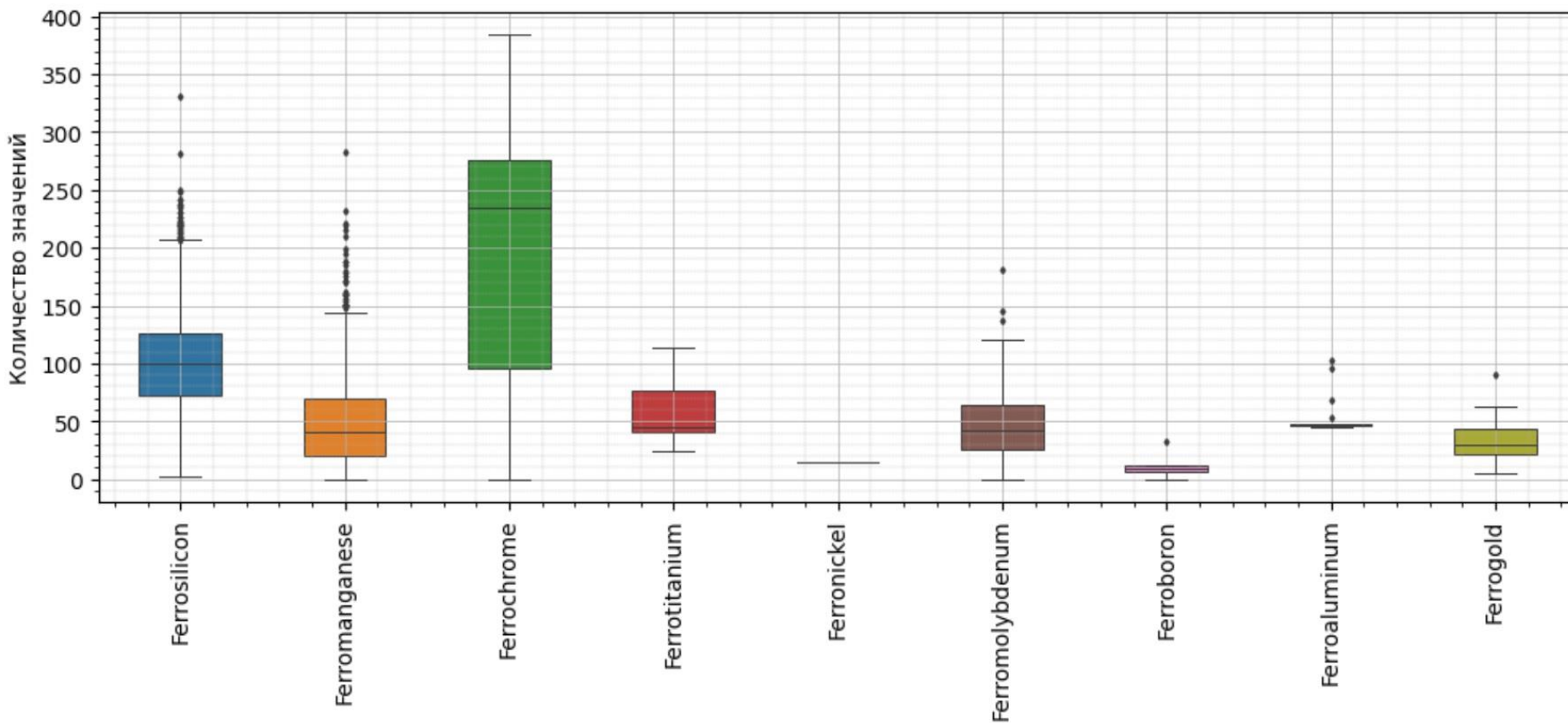


таблица содержит время добавления того или иного сыпучего материала в сплав, т.е. по всем параметрам является чисто информационной и для Machine Learning необязательной.



# данные о проволочных материалах (объём)



- изучили датасет
- определили природу пропусков
- отметили как потенциальные к удалению те признаки, в которых количество пропусков превышает 98%
- пропуски заполнили нулём

# данные о продувке сплава аргоном

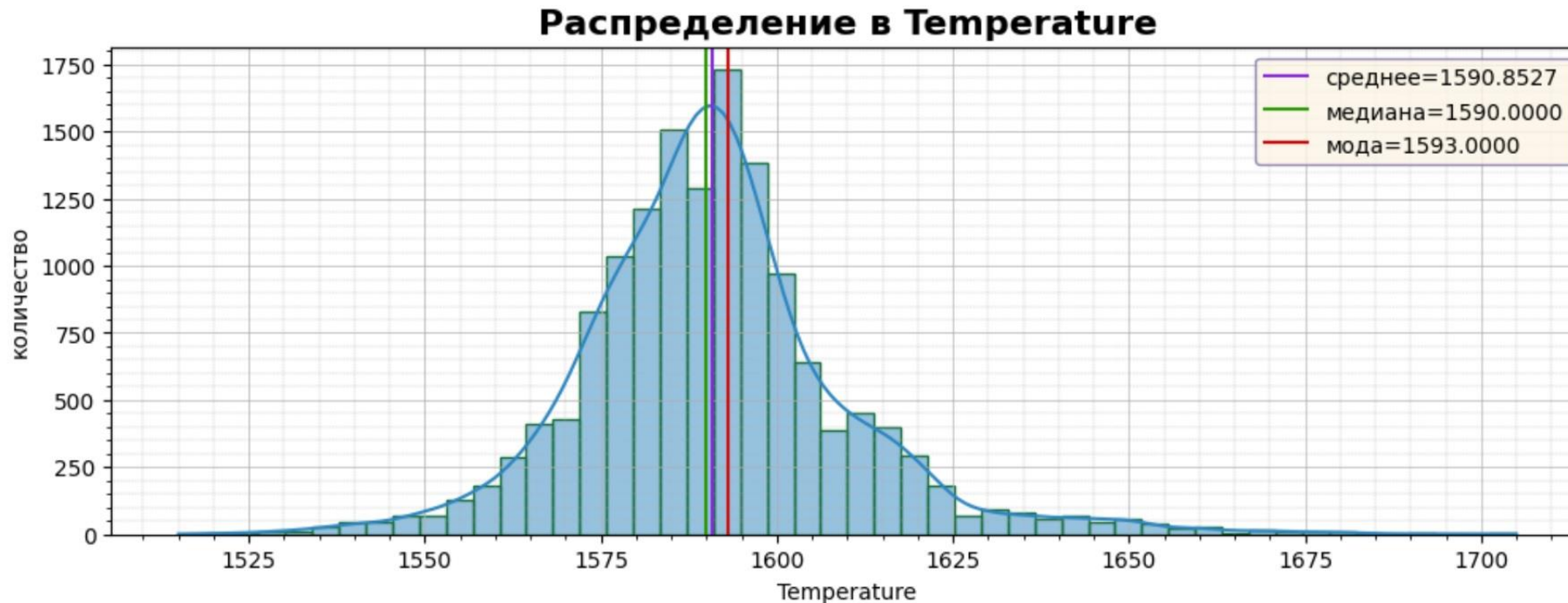
распределение с шагом в 10% для Ar



- изучили датасет
- выявили аномалию в значениях, находящихся выше 90% всех остальных данных
- выбросы удалили.



# данные об измерениях температуры



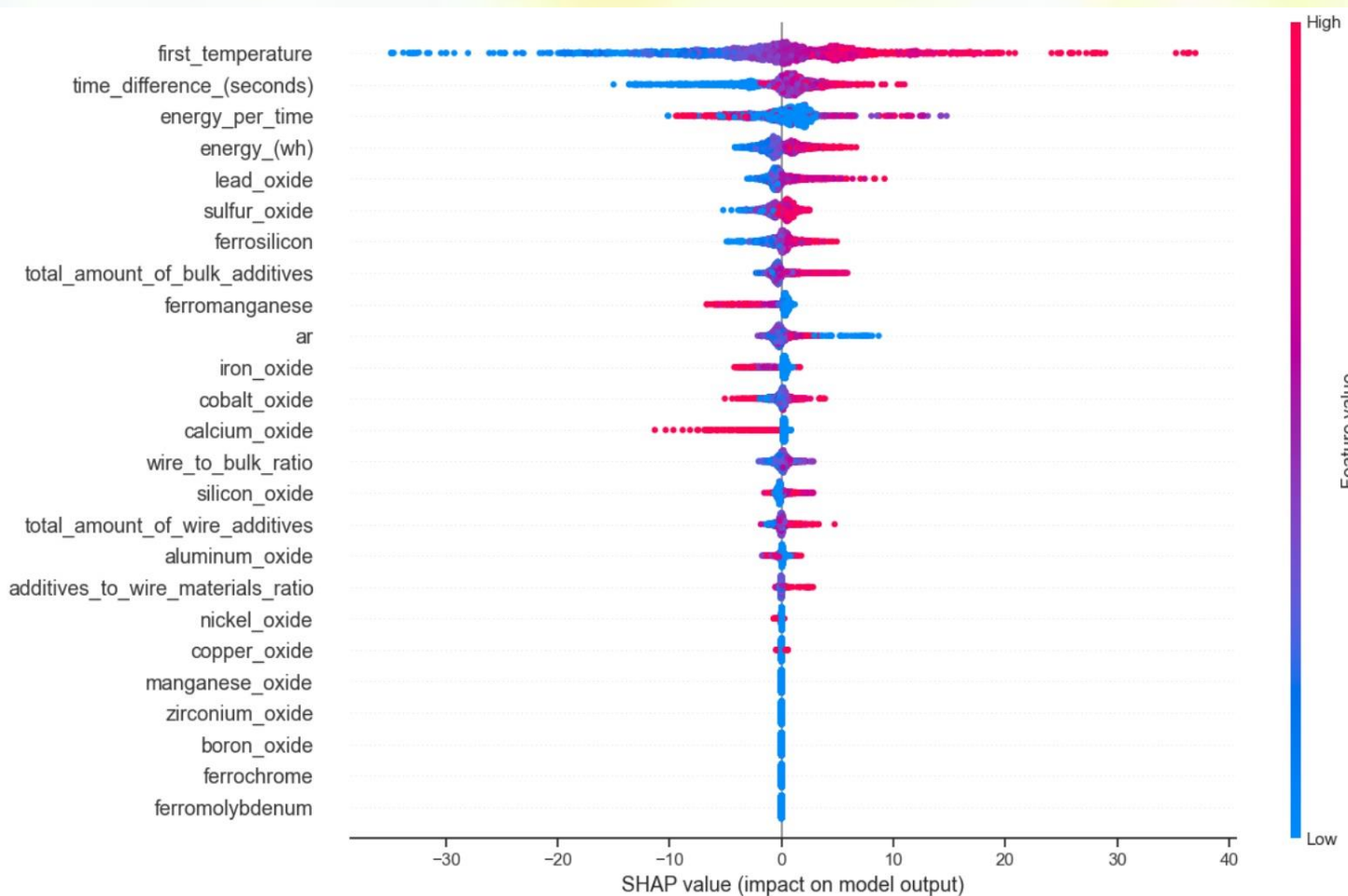
- изучили датасет
- исправили формат дат
- отсортировали данные по дате
- удалили аномально низкие температуры стали
- удалили все промежуточные температуры, оставив первую и последнюю и замерив в секундах время между ними

# feature engineering

- **total\_amount\_of\_bulk\_additives** - суммарное количество сыпучих добавок
- **average\_amount\_of\_bulk\_additives** - среднее количество сыпучих добавок
- **total\_amount\_of\_wire\_additives** - суммарное количество проволочных добавок
- **average\_amount\_of\_wire\_additives** - среднее количество проволочных добавок
- **total\_additives** - общее суммарное количество всех добавок
- **average\_additives** - общее среднее количество всех добавок
- **wire\_to\_bulk\_ratio** - отношение суммы проволочных материалов к сумме сыпучих материалов
- **additives\_to\_wire\_materials\_ratio** - соотношение всех добавок к проволочным материалам
- **energy\_per\_time** - отношение энергии ко времени



# оценка влияния признаков на baseline-модель



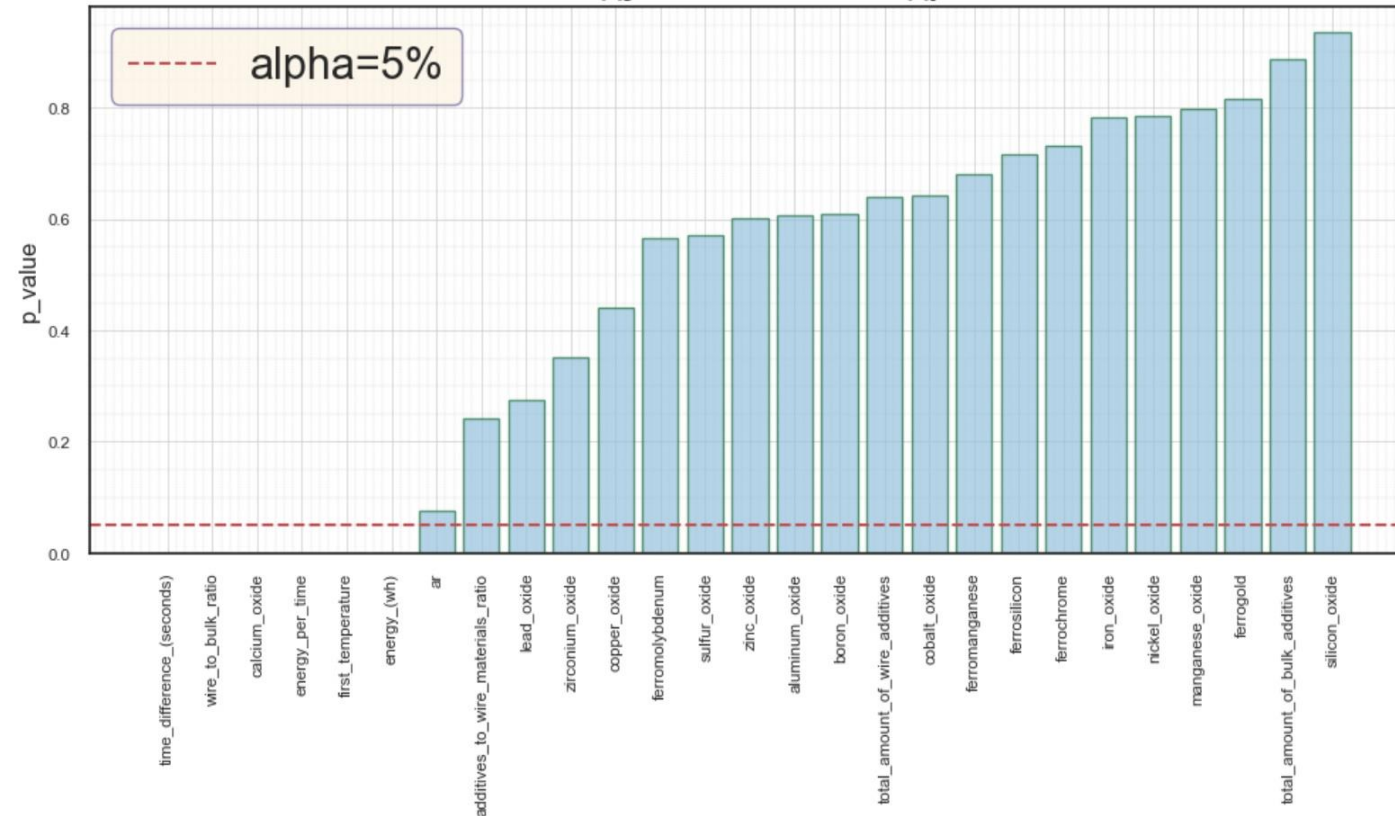
с точки зрения SHAP-модели влияние на целевую переменную имеет всего 21 признак

а вот признаки, которые на модель не влияют, в том числе с т.зр. корреляции Пирсона:

- manganese\_oxide - оксид марганца
- zirconium\_oxide - оксид циркония
- boron\_oxide - оксид бора
- zinc\_oxide - оксид цинка
- ferrochrome - феррохром
- ferrogold - феррозолото

# оценка влияния признаков на baseline-модель

статистический уровень значимости признаков  
по методу наименьших квадратов



по методу наименьших квадратов мы можем отклонить нулевую гипотезу только для шести признаков - это:

- time\_difference\_(seconds) - время легирования
- wire\_to\_bulk\_ratio - отношение проволочных материалов к сыпучим
- calcium\_oxide - оксид кальция
- energy\_per\_time - отношение энергии ко времени легирования
- first\_temperature - первый замер температуры
- energy\_(wh) - затраты энергии на арковую плавку в ватт-часах

по остальным признакам мы не можем отклонить нулевую гипотезу, а это значит, что с вероятностью в 95% они являются статистически незначимыми и, скорее всего, не влияют на целевую переменную.



# => для Machine Learning остаются следующие признаки:

- **time\_difference\_(seconds)** - время легирования
- **first\_temperature** - первый замер температуры
- **energy\_(wh)** - затраты энергии на арковую плавку в ватт-часах
- **calcium\_oxide** - оксид кальция
- **luminum\_oxide** - оксид алюминия
- **silicon\_oxide** - оксид кремния
- **iron\_oxide** - оксид железа
- **nickel\_oxide** - оксид никеля
- **copper\_oxide** - оксид меди
- **cobalt\_oxide** - оксид кобальта
- **lead\_oxide** - оксид свинца
- **sulfur\_oxide** - оксид серы
- **ferrosilicon** – ферросилиций
- **ferromanganese** – ферромарганец
- **ferromolybdenum** – ферромолибден
- **ar** - аргон для продуве смеси
- **total\_amount\_of\_bulk\_additives** - суммарное количество сыпучих добавок
- **total\_amount\_of\_wire\_additives** - суммарное количество проволоочных материалов
- **wire\_to\_bulk\_ratio** - отношение проволоочных материалов к сыпучим
- **additives\_to\_wire\_materials\_ratio** - соотношение всех добавок к проволоочным
- **energy\_per\_time** - отношение энергии ко времени легирования

# МЫ ПОДГОТОВИЛИ ДАННЫЕ К Machine Learning:

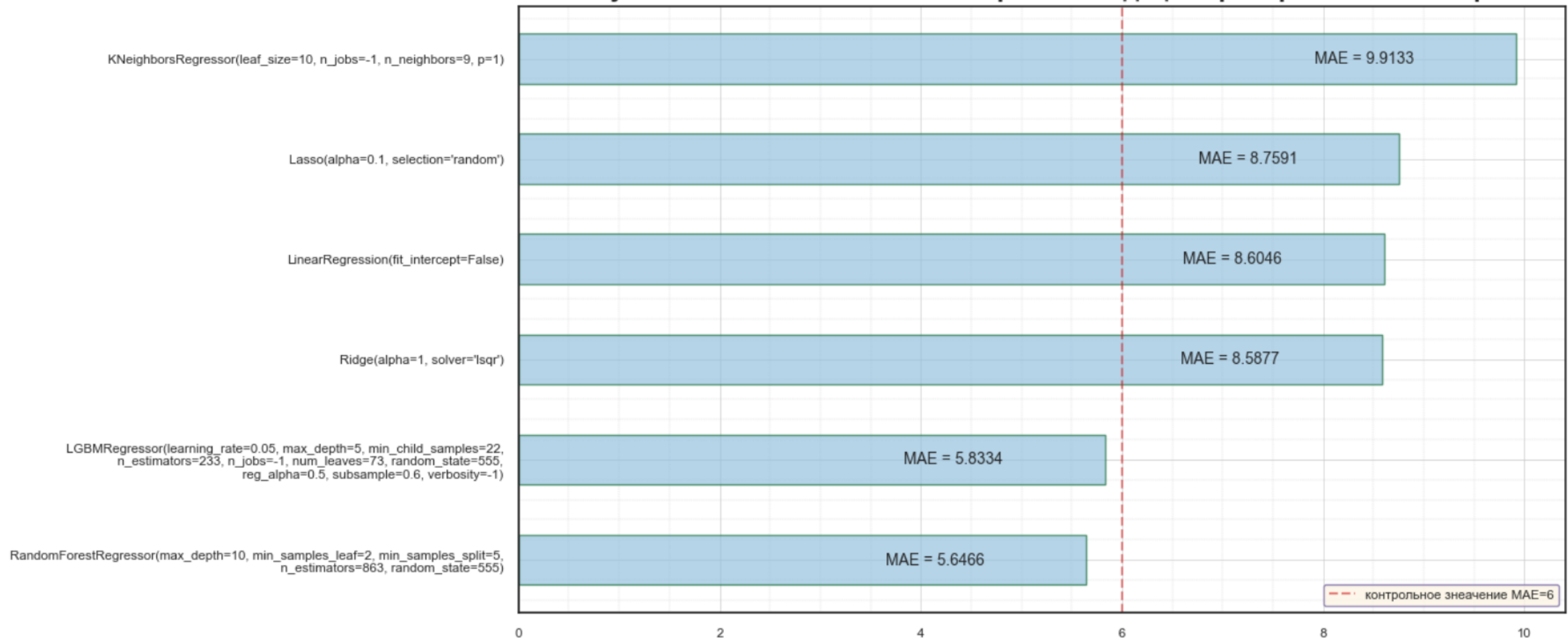
- объединили все имеющиеся в нашем распоряжении данные в единый датасет
- определили целевую переменную - это *last\_temperature* время последнего замера температуры
- feature engineering - сгенерировали ряд новых признаков
- разделили датасет на две выборки - тренировочную и тестовую, оставили на тест 25% данных
- при помощи нескольких методов, в т.ч. с оценкой baseline-модели изучили влияние признаков на целевую переменную. Эти методы:
  - оценка корреляции Пирсона
  - интерпретация SHAP-модели
  - статистическая оценка модели регрессии методом наименьших квадратов
- определились с признаками, которые будут участвовать в ML-обучении, лишние удалили, нужные оставили





# Поиск оптимального алгоритма и создание ML-модели

Результаты по значению MAE на кросс-валидации тренировочной выборки



# создали модель и проверили её на отложенной выборке

- проанализировали сводную таблицу с полученными предсказаниями от разных алгоритмов
- в цикле попарно перебрали все алгоритмы и посчитаем MAE на их усреднённых предсказаниях
- выбрали алгоритм, показавший лучшее значение MAE, им оказался:

**RandomForestRegressor + LGBMRegressor**

**MAE = 5.064361**

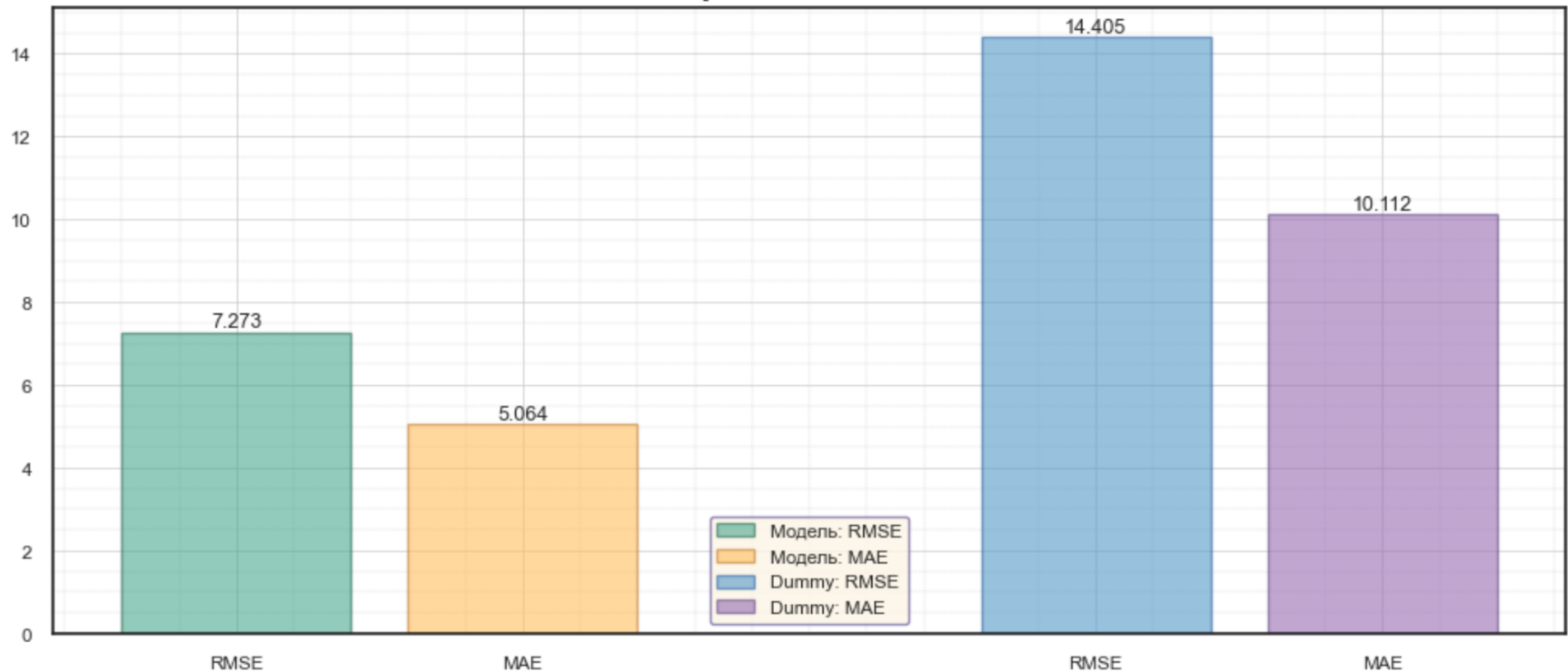
**RMSE = 7.273358**





# проверка модели на адекватность

Сравнение предсказаний модели и dummy-модели,  
стратегия: mean



# что могло помешать исследованию и повлиять на качество разработанной модели

- **отсутствие информации по энергопотреблению, ради потенциальной экономии которого мы и формировали модель** - значения по энергопотреблению в качестве дополнительной бизнес-метрики позволили бы нам лучше отстроить модель, так как при её калибровке мы бы ориентировались в том числе и на расход электричества. Кроме того, мы смогли бы явно показать бизнес-пользу при экономии электричества, показав разницу в его расходе с использованием модели и без неё
- **недостаток технологических данных** - лишней информации не бывает и для точности предсказания модели не помешали бы дополнительные сведения, касающиеся процесса легирования стали, к примеру: данные о химическом составе исходного сплава, данные о скорости перемешивания сплава во время легирования, данные о физических свойствах добавляемых материалов, таких как плотность, теплоемкость и теплопроводность, технические характеристики самого ковша (объём, толщина огнеупорного кирпича, скорость подачи тока и стали в ковш и др.)
- **ограниченность информации по времени** - в нашем распоряжении были только данные с марта по июнь 2019-го года, данные за больший срок могли бы улучшить точность предсказания
- **природа аномалий** - некоторые показатели признаков "выбивались" из общей картины, необходимо уточнение у технических специалистов металлургического комбината природы подобных показателей и, если они являются нормой и частью производственного процесса, а не какой-то механической ошибкой, то их необходимо добавить к модели, дообучив её





использование разработанной модели машинного обучения для оптимизации производственных расходов металлургического комбината:

- имеет значительный потенциал для снижения затрат
- напрямую влияет на повышение качества продукции
- улучшает производственный процесс
- снижает износ оборудования
- в процессе эксплуатации может найти скрытые паттерны и тенденции, влияющие на производство
- влияет на инновационные преимущества комбината
- делает производство экологически более безопасным
- положительно влияет на репутацию и конкурентоспособность комбината



внедрение модели, позволяющий экономить  
энергозатраты на производстве, станет важным шагом  
в направлении устойчивого развития комбината и  
повышения эффективности в производственной цикле.

