**Министерство науки и высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**“Национальный исследовательский университет ИТМО”**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

­­

**Использование машинного обучения для выбора оптимальных параметров модели СЖИГАНИЯ ГАЗОВОГО ТОПЛИВА В ПАРОГЕНЕРАТОРЕ**

Автор\_\_\_Дятлов Михаил Александрович\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, Имя, Отчество)

Направление подготовки (специальность)\_\_\_\_\_11.03.02\_\_\_\_\_\_\_

(код, наименование)

\_\_\_\_\_\_Инфокоммуникационные технологии и системы связи \_\_\_\_\_\_

Квалификация \_\_\_\_бакалавр\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(бакалавр, магистр, инженер)\*

Руководитель ВКР \_\_\_Ананченко И.В. к.т.н.\_\_ \_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, И., О., ученое звание, степень)

Санкт-Петербург, 2020 г.

Обучающийся\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Дятлов Михаил Александрович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО полностью)

Группа\_\_\_\_\_К3420\_\_\_\_\_Факультет \_инфокоммуникационных технологий\_\_\_\_\_\_\_\_\_\_\_\_

Направленность (профиль), специализация \_\_\_\_интеллектуальные инфокоммуникационные системы\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата защиты “\_18\_”\_\_июня\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2020 г.

Секретарь ГЭК \_Златина С.В.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО) (подпись)

**Министерство науки и высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО"**

**УТВЕРЖДАЮ**

Руководитель ОП

\_\_\_Хоружников С.Э.\_\_ \_\_\_\_\_\_\_\_\_\_\_

(Фамилия, И.О.) (подпись)

«\_16\_» «\_\_\_\_января\_\_\_\_\_» 2020\_\_ г.

**ЗАДАНИЕ**

**НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ**

**Обучающийся**\_\_Дятлов Михаил Александрович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО полностью)

**Группа**\_\_\_\_К3420\_\_\_\_\_**Факультет/институт/кластер**\_факультет инфокоммуникационных технологий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Квалификация** \_\_\_\_\_\_\_\_\_бакалавр\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(магистр, инженер, бакалавр)\*\*

**Направление подготовки** \_\_\_11.03.02 Инфокоммуникационные технологии и системы связи

(код, название направления подготовки)

**Направленность (профиль) образовательной программы** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ интеллектуальные инфокоммуникационные системы\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Специализация** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Тема ВКР**\_\_\_\_\_\_\_ использование машинного обучения для выбора оптимальных параметров модели сжигания газового топлива в парогенераторе\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Руководитель**\_\_\_\_\_\_Ананченко Игорь Викторович, университет ИТМО, доцент факультета инфокоммуникационных технологий, кандидат технических наук \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО полностью, место работы, должность, ученая степень, ученое звание)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2 Срок сдачи студентом законченной работы до «**\_\_\_3\_\_» «\_\_\_\_\_июня\_\_\_\_\_\_\_» 2020\_\_\_г.

**3 Техническое задание и исходные данные к работе**

1. Исходные данные - данные, собранные в течении нескольких дней на котельной.
2. Построить модель регрессии расхожа газа и расхода воды.
3. Скорректировать модель.
4. На основе матрицы корреляции, используя новые параметры и решить задачу поиска управляющих параметров в зависимости от текущего значения пара в барабане котла, минимизируя отношение расхода газа к расходу воды.

**4 Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов)**

1. Работа с реальными данными с датчиков котельной и построение модели регрессии для расхода газа и расхода воды (возможно концентрации СО и О2).
2. Улучшение модели путем подбора новых параметров мониторинга и добавления их в матрицу признаков.
3. При помощи обученных моделей предсказывать значения расхода газа и расхода воды для измененных признаков. Находить значения признаков, при которых отношение расхода газа к расходу воды минимальное, при этом давление в барабане неизменно.

**5 Перечень графического материала (с указанием обязательного материала**)

1. Схема решения задачи методами анализа данных.
2. Принципиальная технологическая схема барабанного котла.
3. Зависимости выбросов и КПД от коэффициента избытка воздуха.
4. Принципиальная схема одной из автоматизированных систем управления качеством сжигания топлива в котлоагрегате.
5. Программная реализация 1.1.
6. Программная реализация 1.2.
7. Программная реализация 1.3.
8. Программная реализация 2.1.
9. Программная реализация 2.2.
10. Программная реализация 2.3.
11. Программная реализация 3.1.
12. Программная реализация 3.2.
13. Программная реализация 3.3.
14. Программная реализация 3.4.
15. Программная реализация 3.4.
16. Программная реализация 3.6.
17. Программная реализация 3.7.
18. График отношения Fг/Fв, ИМП и ИМГ

**6 Исходные материалы и пособия**

нет\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**7 Дата выдачи задания «**\_16\_» «\_\_\_января\_\_\_» 2020\_\_г.

Руководитель ВКР\_\_ Ананченко И.В.\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись)

**Министерство науки и высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО"**

**АнНотация**

**ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ**

**Обучающийся**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Дятлов Михаил Александрович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО)

**Наименование темы ВКР:**\_\_\_\_\_\_\_\_ использование машинного обучения для выбора\_\_\_\_\_\_

оптимальных параметров модели сжигания газового топлива в парогенераторе\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Наименование организации, где выполнена ВКР**\_\_\_\_\_\_АО НПФ «УРАН-СПб»\_\_\_\_\_\_\_\_\_\_

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ**

1 Цель исследования\_\_\_\_ изучение возможностей улучшения модели сжигания газового топлива в парогенераторе методами машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2 Задачи, решаемые в ВКР \_\_оптимизация параметров модели сжигания газового топлива\_\_

3 Число источников, использованных при составлении обзора\_4\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4 Полное число источников, использованных в работе \_4\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5 В том числе источников по годам

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Отечественных** | | | | **Иностранных** | | |
| Последние 5 лет | От  5 до 10 лет | Более  10 лет | Последние  5 лет | | От  5 до 10 лет | Более  10 лет |
| 2 | 2 |  |  | |  |  |

6 Использование информационных ресурсов Internet\_да, 2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Да, нет, число ссылок в списке литературы)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7 Использование современных пакетов компьютерных программ и технологий (Указать, какие именно, и в каком разделе работы)

|  |  |
| --- | --- |
| **Пакеты компьютерных программ и технологий** | **Раздел работы** |
| Python | 2 |
| Scikit-learn | 2 |
| Pandas | 2 |
|  |  |
|  |  |

8 Краткая характеристика полученных результатов \_по результатам исследования найдены параметры процесса сжигания газового топлива в парогенераторе, изменение которых приведет к наиболее существенным изменения эффективности работы котлоагрегата.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

9 Полученные гранты, при выполнении работы \_\_нет\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Название гранта)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

10 Наличие публикаций и выступлений на конференциях по теме выпускной работы\_\_\_нет\_

(Да, нет)

а) 1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Библиографическое описание публикаций)

2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

б) 1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Библиографическое описание выступлений на конференциях)

2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**СОДЕРЖАНИЕ**

[Введение 6](#_Toc41326415)

[1 Анализ предметной области 8](#_Toc41326416)

[1.1 Анализ данных 8](#_Toc41326417)

[1.2 Обработка первоначальных данных 9](#_Toc41326418)

[1.3 Алгоритмы машинного обучения 9](#_Toc41326419)

[1.3.1 Регрессионный анализ 10](#_Toc41326420)

[1.3.2 Линейная регрессия 11](#_Toc41326421)

[1.3.3 Метрики качества регрессии 12](#_Toc41326422)

[1.4 Композиции алгоритмов 14](#_Toc41326423)

[1.5 Программная реализация 16](#_Toc41326424)

[1.6 Описание объекта 17](#_Toc41326425)

[1.7 Оптимизируемый процесс 18](#_Toc41326426)

[1.8 Автоматизация регулирования 19](#_Toc41326427)

[2 Основная часть 21](#_Toc41326428)

[2.1 Работа с реальными данными 21](#_Toc41326429)

[2.1.1 Задача 21](#_Toc41326430)

[2.1.2 Представление данных 21](#_Toc41326431)

[2.1.3 Программная реализация 23](#_Toc41326432)

[2.1.4 Сравнение алгоритмов 25](#_Toc41326433)

[2.1.5 Анализ 26](#_Toc41326434)

[2.1.6 Вывод 27](#_Toc41326435)

[2.2 Улучшение модели 28](#_Toc41326436)

[2.2.1 Задача 28](#_Toc41326437)

[2.2.2 Подбор новых признаков 28](#_Toc41326438)

[2.2.3 Выбранные признаки. 29](#_Toc41326439)

[2.2.4 Программная реализация. 29](#_Toc41326440)

[2.2.5 Новые показатели моделей 31](#_Toc41326441)

[2.2.6 Анализ 31](#_Toc41326442)

[2.2.7 Вывод 32](#_Toc41326443)

[2.3 Минимизация расхода газа 32](#_Toc41326444)

[2.3.1 Задача 32](#_Toc41326445)

[2.3.2 Анализ возможного решения 32](#_Toc41326446)

[2.3.3 Программная реализация 33](#_Toc41326447)

[2.3.4 Принципиальное представление 36](#_Toc41326448)

[2.3.5 Нахождение значений параметров для минимизации 37](#_Toc41326449)

[2.3.6 Анализ 39](#_Toc41326450)

[2.3.7 Вывод 39](#_Toc41326451)

[Заключение 40](#_Toc41326452)

[Список литературы 41](#_Toc41326453)

Введение

В России 70% электроэнергии производится на тепловых электростанциях и котельных. В связи с этим очень важно использовать органическое топливо в ТЭС и котельных максимально эффективно. Кроме задачи уменьшения потребляемого топлива для экономической эффективности есть не менее важная задача уменьшения вредных выбросов в атмосферу. Энерго-экологическая оптимизации процесса сжигания газового топлива в парогенераторе производится в том числе путем внедрения информационных технологий автоматизированного управления качеством сжигания топлива.

Научно производственная фирма «Уран-СПБ» уже 25 лет решает задачи энерго-экологической оптимизации парогенераторов. На данный момент «Факел-2010» новейшая система фирмы. Научными сотрудниками уже достаточно изучены технические аспекты оптимизации [1]. Так же в 2019 была написана бакалаврская работа на тему оптимизации технологического процесса парогенератора и в ней были применены методы линейного и нелинейного программирования.

Цель этой работы в изучении возможностей улучшения модели сжигания газового топлива в парогенераторе методами машинного обучения. Машинное обучение используется для предсказания параметров котла, благодаря чему можно составить анализ работы котлоагрегата и раньше и эффективнее применять регуляционные механизмы.

Область исследования – информационные технологии в тепловой энергетике.

Объект исследования – процесс сжигания газового топлива в парогенераторе.

Предмет исследования – котлоагрегат котельной в общем виде.

Теоретическая и методологическая основа исследования состоит из уже зарекомендовавших себя методов машинного обучения и общих принципов работы котлоагрегата.

Информационная база исследования – научные статьи, представленные фирмой «Уран-СПБ», о возможностях оптимизации процессов котельной и энерго-экологических проблемах эффективности и базы знаний о методах анализа данных в сети интернет.

Научная новизна исследования в применение машинного обучения для последующего анализа. В большинстве случаев задача решается путем обучения модели и последующим предсказанием на данных, которое дает искомый ответ. В этой работе модели использовались для предсказания на данных, параметры которых были вручную изменены, для поиска оптимальных значений этих параметров.

Практически результаты исследования помогут развитию автоматизированных систем управления качеством сжигания топлива в котлоагрегате.

По результатам исследования найдены параметры процесса сжигания газового топлива в парогенераторе, изменение которых приведет к наиболее существенным изменения эффективности работы котлоагрегата.

Работа состоит из теоретической части (12 страниц) и основной части. Основная часть разбита на 3 подраздела по 5-7 страниц, в каждом подразделе есть своя задача, описание решения задачи, программная реализация, анализ и вывод.

# Анализ предметной области

## Анализ данных

Анализ данных – это совокупность методов и алгоритмов из области математики и информатики для обработки экспериментальных данных и представления их в оптимальном виде для извлечения ценной информации и принятия решений.

Алгоритмы машинного обучения на данный момент плохо работают с просто пластами произвольных данных, поэтому перед построением алгоритма, решающего задачу, нужно обработать данные и представить их в оптимальном виде. Такая предобработка называется препроцессингом (рис. 1).

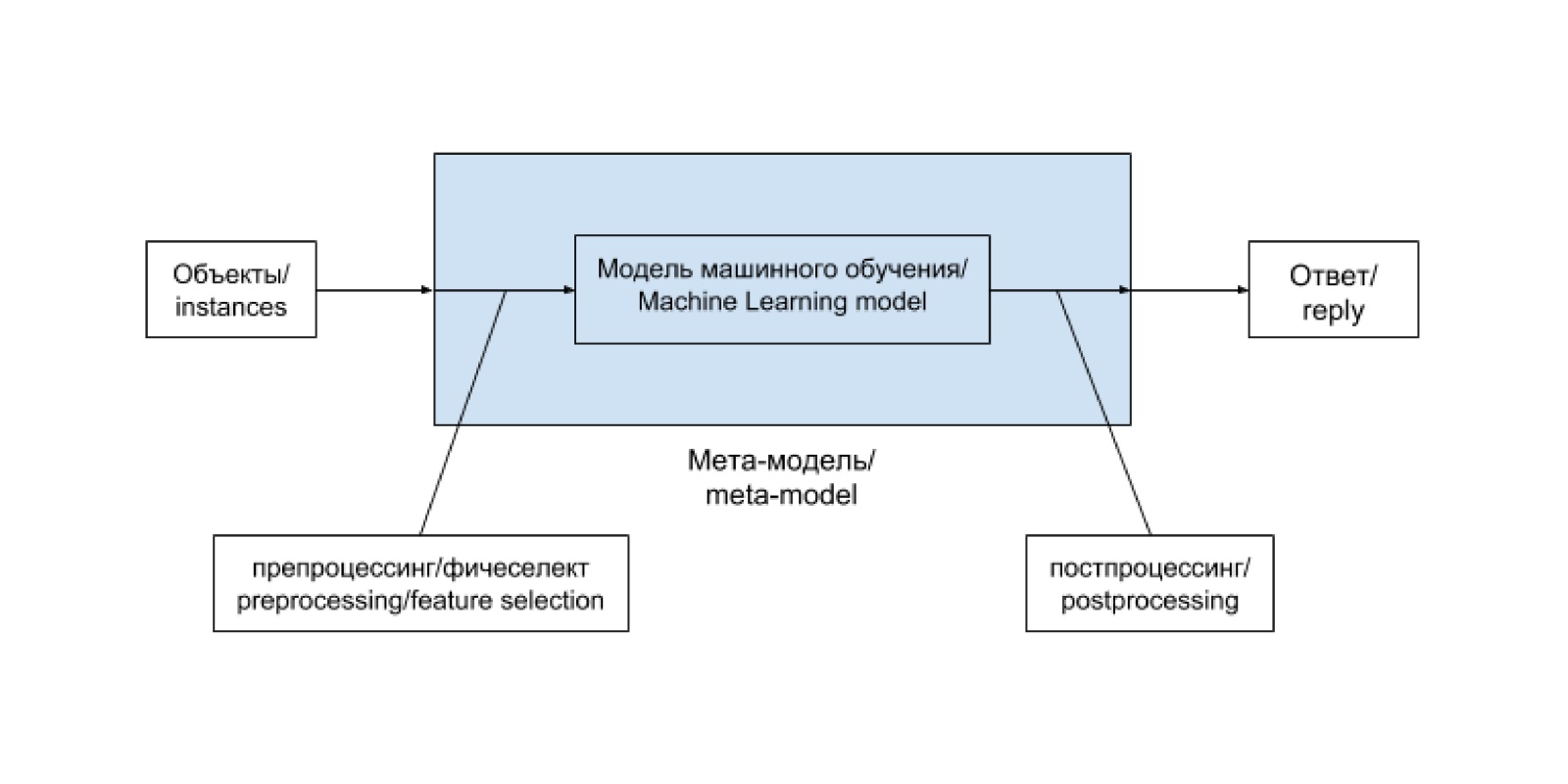


Схема решения задачи методами анализа данных.

На этапе препроцессинга выборка разбивается на обучающую, по которой модель обучается, и контрольную (или тестовую), на которой проверяется правильность предсказаний модели. Далее от изначальной выборки чаще всего бралось 75% на обучающую и 25% на контрольную, если не указано обратного.

## Обработка первоначальных данных

В ходе обучения модели могут возникнуть такие распространенные проблемы как: мультиколлинеарность и переобучение.

Мультиколлинеарность – это линейная зависимость между входными параметрами модели. При этом модель становится не надежной.

Переобучение – это явление, когда модель подстраивается под значения обучающей выборки, но не действенна для других данных. В этом случае модель почти невозможно применять на новых данных, что лишает ее какого-либо смысла.

Эти проблемы можно решать, например, методами регуляризации или методом главных компонент.

L1-регуляризация – обнуляет веса неинформативных признаков.

L2-регуляризация – сокращает веса линейно зависимых признаков, но обычно не обнуляет их.

Метод главных компонент (МГК) – оставляет наиболее значимые признаки.

## Алгоритмы машинного обучения

Алгоритмы машинного обучения – это методы решения задачи путем обучения на большом множестве схожих задач. Оно применяется в большом количестве прикладных задач в различных областях, где есть большие объемы собранных данных и требуется решать задачи автоматизации принятия решений, предсказаний и так далее.

Алгоритм машинного обучения – это недетерминированный алгоритм, от обученной модели мы ожидаем верный ответ, но не можем его требовать. Для алгоритмов машинного обучения результат не предопределен, и любая модель иногда ошибается на каких-то специфичных вариантах входных данных.

Виды моделей машинного обучения:

* + - обучение с учителем
      * классификация
      * регрессия
      * ранжирование
    - обучение без учителя
* кластеризация
* уменьшение размерности
* обучение с частичным привлечением учителя
* обучение с подкреплением

В данной работе используется метод регрессии.

### Регрессионный анализ

Регрессия – зависимость отражающая связь одной переменной от независимых между собой переменных.

Регрессионная модель – это функция, построенная методами регрессионного анализа, которая приближена к исходной неизвестной функции зависимостей.

Регрессионный анализ – решает задачу предсказания количественной целевой переменной на основе векторов признаков (других переменных). Если есть *N* строк данных, состоящих из параметров, среди которых выделяется один целевой, регрессионный анализ предсказывает какое значение целевой переменной будет в *N*+1-ой строке, если известны остальные признаки *N*+1-ой строки [2].

Приближение регрессионной модели к исходной функции осуществляется путем изменения весов регрессионной модели и нахождения наиболее вероятных значений параметров. Регрессионная модель считается построенной, когда значения её весов зафиксированы. Веса модели – это внутренние параметры модели, каждый вес определяет влияние признака на целевую переменную.

Регрессия делится на линейную и нелинейную. При линейной регрессии гипотеза происхождения данных предполагает линейную зависимость целевой переменной от признаков. Соответственно модель линейной регрессии можно представить в виде скалярного произведения весов модели на признаки. При решении задачи методом нелинейной регрессии предполагается, что зависимость от признаков не линейна.

### Линейная регрессия

Линейная регрессия – самый распространенный алгоритм регрессионного анализа в машинном обучении. Прим­­­еняется при линейной зависимости целевого параметра от признаков. Линейная регрессия парная или простейшая, если входной параметр один и многомерная, если таких параметров множество.

Пусть у нас имеется *n* числовых признаков в выборке и­­­з *m* объектов *f*1(*x*),…,f*n*(x). По этой обучающей выборке нужно построить модель многомерной линейной регрессии. Результат – это взвешенная сумма значений признаков (весовые коэффициенты). Формула получившийся функции:

Матричное представление:

F*m×n* – матрица объекты-признаки,

*ym×1* – вектор ответов,

α*m×1* – вектор коэффициентов.

При обучении модели нужно корректировать модель с каждой ошибкой. Для уменьшения ошибок модели чаще всего используется метод наименьших квадратов, т.к. в нем происходит достаточно сильное наказание за ошибку.

Задача минимизации функционала ошибки методом наименьших квадратов для линейной регрессии:

Задача линейной регрессии состоит в том, чтобы найти вектор коэффициентов α.

### Метрики качества регрессии

Для определения того, насколько хорошо работает модель нужна метрика качества. Метрика качества предсказания алгоритма – это оценка качества предсказания в сравнении с реальными данными (ответами). При решении каждой задачи может применяться различная метрика качества, в зависимости от того, какая особенность или характеристика предсказываемого объекта или предсказания наиболее существенна в рамках самой задачи и сферы ее применения. Основные симметричные метрик качества для регрессии:

* средняя абсолютная ошибка,
* средняя процентная ошибка,
* среднеквадратичная ошибка,
* корень из среднеквадратичной ошибки,
* коэффициент детерминации.

Средняя абсолютная ошибка – эта метрика измеряет среднее значение отклонения предсказания от реального значения по абсолютной величине, то есть непосредственно по величине параметра. Средняя абсолютная ошибка вычисляется по формуле:

Где α(*xi*) – значение, предсказанное моделью на *i*-ом объекте, *yi* – реальное значение целевого параметра объекта выборки.

Средняя процентная ошибка – показывает величину отклонения относительно действительного значения в процентах.

Формула средней процентной ошибки

Среднеквадратичная ошибка – вычисляет средний квадрат отклонения предсказания от *y*. Преимущество над средней абсолютной ошибкой в том, что ошибки гораздо сильнее влияют на показатель ошибки, это полезно, если важна точность модели.

Формула среднеквадратичной ошибки:

Корень из среднеквадратичной ошибки – вычисляется среднеквадратичная ошибка, а потом берется корень получившегося значения. Так влияние каждой отдельной ошибки все еще существенно, но единица измерения остается такой же как у ответов.

Формула корня из среднеквадратичной ошибки:

Коэффициент детерминации – это доля разброса значений целевой переменной.

Формула коэффициента детерминации:

ȳ - среднеарифметическое значение целевого параметра по всем объектам.

Коэффициент детерминации может принимать значения:

* 0 ≤ R2 ≤ 1 – для разумных моделей
* R2 = 1 – идеальная модель
* R2 = 0 – модель на уровне константной (константная модель – выдает всегда среднее значение целевой переменной в обучающей выборке)
* R2 < 0 – модель хуже константной

Приемлемой для линейной регрессии с константной считается модель с коэффициентом дисперсии больше 0,5. При коэффициенте детерминации больше 0,8 качество модели можно считать хорошим. Значение в единицу означает, что модель в точности описывает исходную зависимость.

## Композиции алгоритмов

При разбиении выборки для множества алгоритмов, для улучшения качество композиции применяются такие методы:

* Разбиение выборки на случайные подвыборки.
* Случайное изменение весов объектов.
* Выделение случайных подмножеств признаков для базовых алгоритмов.
* Применение различных видов базовых алгоритмов.

Бэггинг – обучает базовые алгоритмы на подвыборках, длина которых равна длине основной выборки, но состоит она из неполного множества объектов первоначальной выборки, в которой существуют повторения объектов [3, 4].

Метод случайных подпространств – базовые алгоритмы обучаются по случайным подмножествам признаков, число признаков определяется заранее и одинаково для каждого алгоритма.

Метод случайных подпространств и бэггинг можно объединить в один метод – случайный лес.

Случайный лес – бэггинг над решающими деревьями. Решающее дерево – дерево из вопросов, в котором на вершинах расположены условия а на листах прогнозы.

Простое голосование – композиция классификаторов, при которой объект относится к тому классу, который был выбран большинством классификаторов. Если вероятность правильного определения у каждого классификатора больше 0,5, то точность композиции возрастает с увеличением числа базовых классификаторов в ней.

Взвешенное голосование – отличается от простого голосования тем, что каждый базовый алгоритм имеет собственный весовой коэффициент.

Бустинг – способ построения линейной композиции базовых алгоритмов, при котором каждый следующий базовый алгоритм настраивается так, чтобы покрывать слабые стороны предыдущих алгоритмов.

И бэггинг, и бустинг строят композиции из уже известных базовых алгоритмов, но в отличии от бэггинга бустинг:

1. Позволяет применять произвольную функцию потерь, следовательно может быть применен не только для задач классификации, но и для задач регрессии.
2. Использует взвешенное голосования.

Градиентный бустинг – наиболее популярный алгоритм, включающий в себя функционал множества предшествующих алгоритмов бустинга.

Функция потерь – это функция, характеризующая потери при не верном ответе, является мерой расхождения между правильным ответом и предсказанием.

Пусть существует функция потерь *L*(*b*, *y*), где *b* – оценка, которую дает композиция, а *y* – правильный ответ. Тогда задача минимизации функционала ошибки для градиентного бустинга выглядит как

Где T – количество базовых алгоритмов в композиции.

При решении задачи минимизации функционала ошибки алгоритм идет последовательно по каждому базовому алгоритму. При каждом новом шаге алгоритм градиентного бустинга подбирает новый базовый алгоритм и коэффициент так, чтобы приблизиться к антиградиенту функции потерь.

## Программная реализация

Используемый язык программирования для реализации: *Python*. Функции работа с данными и реализованные алгоритмы машинного обучения взяты из сторонних библиотек: *NumPy*, *scikit-learn* и *pandas*.

*NumPy* – библиотека фундаментальных научных вычислений, это мощный инструмент для математических операций в языке *python*. Одним из важнейших для дальнейшей работы инструментов является качественные *N*-мерные массивы и простая работа с ними. Так как данные удобнее всего представлять в матричном виде, *NumPy* важное средство для работы в сфере анализа данных в языке программирования python.

*Pandas* – это библиотека, которая использую функционал библиотеки *NumPy* упрощает работу с массивами данных. Благодаря библиотеке pandas можно легко преобразовать данные из множества форматов, таких как: «csv», «xls» и т.д., в матрицы из библиотеки *NumPy*.

*Scikit-learn* – основная библиотека для работы с машинным обучением в языке python. Библиотека с открытым исходным кодом. *Scikit-learn* содержит огромное количество реализованных методов машинного обучения. От классов различных видов моделей регрессии, классификации и кластеризации до методов их оценивания и методов пре- и постпроцессинга.

Основные методы и классы, использующиеся в работе:

*sklearn.linear\_model.LinearRegression* – Класс реализованной линейной регрессии, обучающуюся с применением метода наименьших квадратов. Как и все алгоритмы машинного обучения в библиотеке *scikit-learn* обучается вызовом метода *fit* с переданными в параметры матрицей признаков *X* и матрицей ответов *y*. Также имеет метод *score*, который принимает матрицы признаков и ответов тестовой выборки и возвращает коэффициент детерминации.

*sklearn.linear\_model.ElasticNet* – Такой же класс линейной регрессии, но внутри себя уже имеет L1 и L2 регулизации.

*sklearn.linear\_model.ElasticNetCV* – Представляет собой такой же класс, как «ElasticNet», но с добавление метода кросс-валидации.

*sklearn.ensemble.GradientBoostingRegressor* – Класс, реализующий метод градиентного бустинга над линейной регрессией. В параметрах можно задать (*n\_estimators*) количество шагов бустинга, выбрать функцию потерь и задать другие свойства.

*sklearn.model\_selection.train\_test\_split* – для простого разбиения данных на тестовую и обучающую выборки. В параметрах метода можно задать размер тестовой выборки, также можно выбрать перемешивать данные между собой или нет.

*pandas.read\_excel* – для чтения данных из формата «xls» и объединения столбцов «Дата» и «Время» в один через параметр «parse\_dates».

*pandas.concat* – для объединения датасетов в один.

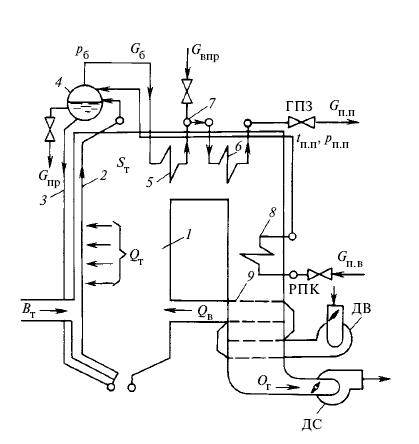
## Описание объекта

Котельная – технологический комплекс, предназначенный для выработки пара и горячей воды. Осуществляется это за счет нагрева воды в котле котельной. Работать котельные могут для обеспечения производственных нужд или для отопления. Для нагрева могут использоваться различные виды топлива: газ, жидкое топливо или твердое топливо. Котельные используются при централизованном тепло- и пароснабжении или при местном теплоснабжении зданий.

Основным устройством в котельной является котел, котлы бывают:

* Паровые
* Водогрейные
* Смешанные
* Диатермические

Котлоагрегат (котел) – обеспечивает нагрев непрерывно поступающей воды путем сжигания органического топлива.



Принципиальная технологическая схема барабанного котла.

ГПЗ – главная паровая задвижка; РПК – регулирующий питательный клапан; 1 – топка; 2 – циркуляционный контур; 3 – опускные грубы; 4 – барабан; 5, 6 – пароперегреватели; 7 – пароохладитель; 8 – экономайзер; 9 – воздухоподогреватель.

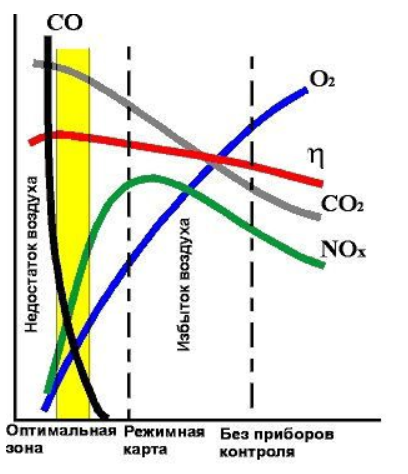
Сжигание топлива производится в топке. Жар от топлива и образующегося дыма нагревает воду.

## Оптимизируемый процесс

Основная задача – энерго-экологическая оптимизация процесса работы котлоагрегата.

Энергетическая оптимизация – по сути задача оптимизации кпд. Основной ресурс котельной – это топливо. При увеличении кпд, меньше топлива расходуется в пустоту. Основное явление, снижающее кпд то химический недожог. Химический недожог регулируется коэффициентом избытка воздуха.

Экологическая оптимизация – уменьшение вредных выбросов котлоагрегата.

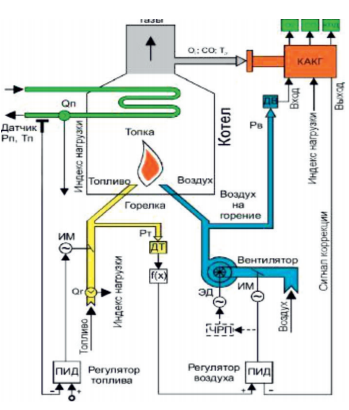


Зависимости выбросов и КПД от коэффициента избытка воздуха.

## Автоматизация регулирования

Для обеспечения лучшей работы котлоагрегата не малую роль играет автоматизация. Большинство котельных работают по режимным картам, которые не учитывают множества динамических факторов. При этом множество котельных завышают количество подаваемого воздуха, для предотвращения химического недожога, чем увеличивают количество вредных выбросов. При этом есть возможность, не теряя энергоэффективность снизить количество воздуха и уменьшить выбросы. Также котлоагрегаты часто работают на ручном регулировании, работники определяют подачу воздуха на глаз, по виду пламени котла. Такое устройство работы приводит к перерасходу топлива и повышенному выбросу загрязняющих веществ в атмосферу.

Контроль подачи воздуха и топлива можно осуществлять автоматизировано. Осуществляется за счет контроля положений исполнительных механизмов подачи топлива и газа и частоты оборотом вентиляторов.



Принципиальная схема одной из автоматизированных систем управления качеством сжигания топлива в котлоагрегате.

Благодаря автоматизации можно:

* Экономить от 6% до 12% топлива.
* Снизить выбросы оксидов азота в атмосферу на 30-40%.
* Снизить энергопотребление вентиляторов на 25-50%.
* Повысить надежность.

# Основная часть

## Работа с реальными данными

### Задача

Работа с реальными данными с датчиков котельной и построение модели регрессии для расхода газа и расхода воды (возможно концентрации СО и О2).

### Представление данных

Реальные данные представляют из себя таблицы замеров с датчиков. Данные полные (нет пропущенных ячеек). Данные взяты за определенный промежуток, в котором каждую минуту собирались данные сс датчиков и записывались в таблицу. Датчики и приборы замеряют более тридцати параметров (таблица 1.1).

Таблица 1.1 Параметры и их обозначения

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Описание параметра | Кр  Обозн | Х  мин | Х  минР | Х  максР | Х  макс |
| Температура воды на входе, гр. Ц | Твв | 0 | 20 | 150 | 180 |
| Температура воды после экономайзера, гр.Ц | Твэ | 0 | 20 | 150 | 180 |
| Давление воды после экономайзера, кг/см2 | Рвэ | 0 | 2 | 5 | 10 |
| Расход воды, м3/ч | Fв | 0 | 2 | 8 | 10 |
| Положение исполнительного механизма воды, % | ИМП | 0 | 0 | 100 | 100 |
| Уровень воды в барабане котла, мм | Lб | 0 | 25 | 75 | 100 |
| Давление пара в барабане, кгс/см2 | Рпб | 0 | 0 | 13 | 16 |
| Расход пара, т/ч | Fп | 0 | 5 | 14 | 15 |
| Температура воздуха на горение, гр.Ц | Тнв | –50 | –40 | 40 | 50 |
| Частота электропитания вентилятора, Гц | fвент | 0 | 0 | 50 | 60 |
| Положение исполнительного механизма воздуха, % | ИМВ | 0 | 0 | 100 | 100 |
| Давление воздуха общее, кПа | Рво | 0 | 5 | 200 | 250 |
| Давление воздуха в горелке 1 (левой), кПа | Рвл | 0 | 5 | 200 | 250 |
| Давление воздуха в горелке 2 (правой), кПа | Рвп | 0 | 5 | 200 | 250 |
| Расход газа, м3/час | Fг | 0 | 300 | 1300 | 2000 |
| Температура газа, гр.Ц | Тг | –20 | –15 | 30 | 40 |
| Положение исполнительного механизма газа, % | ИМГ | 0 | 0 | 100 | 100 |
| Давление газа, общее, кг/см2 | Рго | 0 | 0.15 | 5 | 10 |
| Давление газа, перед горелками, кг/см2 | Ргг | 0 | 0.15 | 5 | 10 |
| Давление газа в горелке 1 (левой), кг/см2 | Ргл | 0 | 0.15 | 2.6 | 5 |
| Давление газа в горелке 2 (правой), кг/см2 | Ргп | 0 | 0.15 | 2.6 | 5 |
| Разрежение в топке котла, Па | Ртк | –250 | –100 | 0 | 250 |
| Содержание оксида углерода, ppm | CO | 0 | 0 | 500 | 2000 |
| Содержание кислорода после котла, % | O2к | 0 | 1.0 | 10 | 15 |
| Содержание кислорода после экономайзера, % | О2э | 0 | 2 | 12 | 15 |
| Разрежение в дымоходе после котла, кПа | Рдк | 0 | 0 | 0.4 | 2.5 |
| Температура дымовых газов после котла, гр. Ц | Тдк | 50 | 100 | 450 | 500 |
| Температура дымовых газов после экономайзера, гр.Ц | Тдэ | 50 | 100 | 350 | 400 |
| Частота электропитания дымососа, Гц | fдым | 0 | 0 | 50 | 60 |
| Положение исполнительного механизма разрежения, % | ИМТ | 0 | 0 | 100 | 100 |
| Разрежение в дымоходе после экономайзера, кПа | Рдэ | 0 | 0 | 0.6 | 2.5 |

Сырые данные представлены в виде таблицы в формате программного обеспечения *Microsoft Excel*.

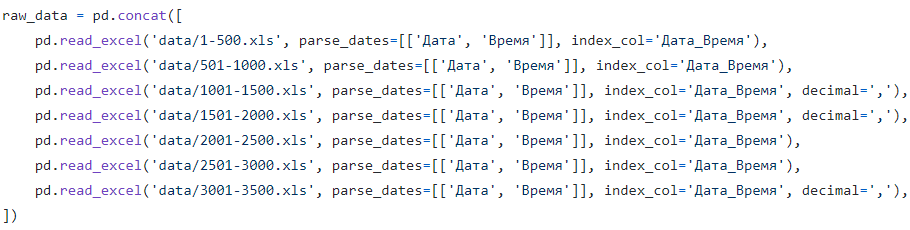
Анализ предстоящего преобразования данных:

1. Данные нужно преобразовать из файлов формата «.xls» к матричному виду.
2. Данные лежат в разных файлах, требуется объединить в один датасет.
3. Нужно выбелить какие столбцы будут целевыми, а какие представляют собой параметры для предсказания.
4. Для каждого выбранного целевого параметра убрать столбцы с прямой корреляцией (применить методы регуляризации).
5. Выборка содержит дату в одном столбце и время в другом, нужно объединить и обеспечить корректную работу с информацией о дате и времени замеров.
6. В некоторых файлах вместо точки как разделителя используется запятая.

В качестве целевых переменных выступают: содержание оксида углерода (CO), содержание кислорода после котла (O2к), расход воды (Fв) и расход газа (Fг). Признаки, выбранные для первого шага: положение исполнительного механизма воды (ИМП), положение исполнительного механизма газа (ИМГ), частота вентилятора, падающего воздух (fвент) и частота вентилятора дымососа (fдым).

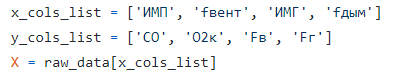
### Программная реализация

Для чтения из файла «.xls» используется метод *read\_excel* а для объединения в один датасет метод *concat*, оба метода из библиотеки pandas. Задать разделитель и объединить дату и время можно с помощью параметров *decimal* и *parse\_dates* метода *read\_excel*.



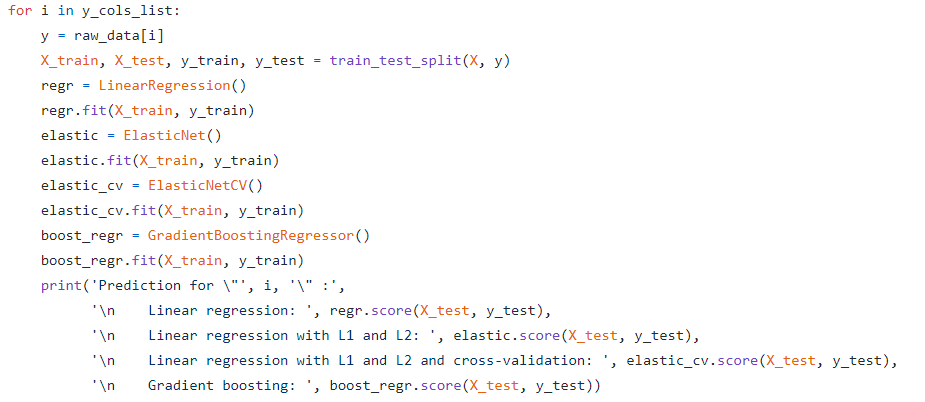
Программная реализация 1.1.

Целевые переменные и признаки разделяются очень просто. В данном случае и для матрицы признаков, и для матрицы ответов были созданы массивы строк с названиями колонок нужных параметров. Сразу строится матрица признаков. Делается это так же, как в массиве происходит обращение по индексу, но вместо индекса передается массив с названиями столбцов. Так из матрицы *raw\_data* библиотека *NumPy* возвращает матрицу признаков.



Программная реализация 1.2.

Оставшаяся часть кода – это цикл по элементам массива *y\_cols\_list*. В этом цикле для каждого элемента массива создается новая матрица ответов из одного столбца. Далее, с помощью уже описанного метода *train\_test\_split*, обе матрицы разбиваются на обучающую и тестовую. Потом объявляются объекты описанных в первой части классов алгоритмов машинного обучения. Каждый объект обучается на обучающей выборке. И на конец в функции print вызываются значения коэффициентов детерминации для переданных в параметры метода score тестовых матриц. Метод *score* сразу предсказывает значения для переданной матрицы признаков, сравнивает их с матрицей ответов и считает, и возвращает коэффициент детерминации. Таким образом для проверки качества модели даже не нужно отдельно вызывать метод предсказания. Далее, когда потребуется предсказывать параметры для новых данных, конечно, будет вызываться метод *predict*. Но на данном этапе важна только проверка качества модели.



Программная реализация 1.3.

### Сравнение алгоритмов

Для выбранных целевых параметров были обучены модели линейной регрессии, линейной регрессии с применением L1 и L2 регуляризации, линейной регрессии с применением L1 и L2 регуляризации и кросс-валидации и градиентного бустинга.

Значения коэффициентов детерминации для различных моделей при различных представлениях выборки представлены в таблице (1.2).

Таблица 1.2 – Сравнение моделей.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | | | |
| Целевая переменная | CO | O2к | Fв | Fг |
| **Не перемешанная выборка** |  |  |  |  |
| Линейная регрессия | 0,11 | 0,14 | 0,87 | 0,67 |
| Линейная регрессия с L1 и L2 регуляризациями | 0,37 | 0,05 | 0,87 | 0,66 |
| Линейная регрессия с L1 и L2 регуляризациями и кросс-валидацией | 0,11 | -0,02 | 0,87 | 0,67 |
| Градиентный бустинг | -0,78 | -0,39 | 0,91 | 0,54 |
| **Перемешанная выборка** |  |  |  |  |
| Линейная регрессия | 0,29 | 0,09 | 0,89 | 0,69 |
| Линейная регрессия с L1 и L2 регуляризациями | 0,30 | 0,18 | 0,89 | 0,68 |
| Линейная регрессия с L1 и L2 регуляризациями и кросс-валидацией | 0,3 | 0,1 | 0,89 | 0,69 |
| Градиентный бустинг | 0,6 | 0,77 | 0,94 | 0,78 |
| **Перемешанная выборка с добавлением столбца «Дата\_Время» в параметры** |  |  |  |  |
| Линейная регрессия | 0,45 | 0,26 | 0,9 | 0,72 |
| Линейная регрессия с L1 и L2 регуляризациями | 0,42 | 0,23 | 0,9 | 0,71 |
| Линейная регрессия с L1 и L2 регуляризациями и кросс-валидацией | 0,01 | 0 | 0 | 0,12 |
| Градиентный бустинг | 0,83 | 0,96 | 0,94 | 0,76 |

­

### Анализ

Из таблицы (1) хорошо видно, как различная работа с данными и применение отличающихся алгоритмов влияет на предсказание содержания кислорода после котла (O2к). Для предсказания этого параметра важным оказывается, что бы данные в выборке располагались не последовательно, а были перемешаны между собой. При этом наиболее эффективным оказался градиентный бустинг.

Но при этом тестовая выборка тоже состояла из замеров в разные моменты времени. Поэтому, для проверки, было решено сначала самостоятельно разбить выборку на две. Потом перемешать обучающую выборку, а тестовую оставить как есть. Суть данной операции в том, что при работе с реальными данными, каждый кусочек данных будет новым, который не находится среди уже известных. Таким образов модели были проверены на состоятельность при будущей работе с новыми данными. Данные представлены в таблице 1.3. Обучающая выборка была также перемешана и столбец «Дата\_Время» включен в признаки.

Таблица 1.3 – Тестовая выборка – последние 1000 строк.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | | | |
| Целевая переменная | CO | O2к | Fв | Fг |
| Линейная регрессия | 0,07 | 0,13 | 0,87 | 0,67 |
| Линейная регрессия с L1 и L2 регулизациями | 0,32 | 0,02 | 0,87 | 0,66 |
| Линейная регрессия с L1 и L2 регулизациями и кросс-валидацией | 0,06 | 0,12 | 0,87 | 0,67 |
| Градиентный бустинг | -1,37 | -0,27 | 0,91 | 0,54 |

Из таблицы 1.3 видно, что при применении нового метода разбиения содержание кислорода и оксида углерода на текущем этапе предсказываются хуже, чем константная модель.

### Вывод

Исходя из первой задачи можно видеть какую большую роль играет работа с данными в машинном обучении. Действительно, линейная регрессия известна достаточно давно, но именно наличие больших объемов данных и вычислительных мощностей сделало машинное обучение возможным в том виде, в котором оно есть сейчас. Поэтому было важно изучить поведение реальных данных, с которыми будет производиться работа.

## Улучшение модели

### Задача

Улучшение модели путем подбора новых параметров мониторинга и добавления их в матрицу признаков.

### Подбор новых признаков

Важным инструментом для выбора признаков является матрица корреляции. Матрицу корреляции можно построит средствами библиотеки pandas, вызвав у матрицы выборки метод corr. В случае с представленными данными с котельной, матрица корреляции уже была построена, поэтому для удобства использоваться будет она.

Матрица корреляции – это матрица или таблица показывающая зависимость переменных между собой. Заголовки столбцов и строк – заголовки параметров, на пересечении находится зависимость между двумя параметрами.

В связи с тем, что количество признаков и целевых переменных не слишком большое, матрица корреляции представлена в удобном виде и процесс местами требует аналитики, новые признаки подбирались вручную.

Сначала для каждого целевого параметра были выбраны параметры со значимым коэффициентом корреляции. Так как эти параметры наиболее значимы и имеют большее влияние на целевую переменную. Далее среди выбранных признаков были найдены признаки, корреляция между которыми превышает значение 0,9. Если два признака сильно коррелируют между собой, выбирался тот признак, который больше влиял на целевую переменную.

Таким образом решалась задача оптимизации модели. Данный метод строится на основном свойстве линейной регрессии, описанном выше. Признаки должны коррелировать с целевой переменной, так как считается, что она зависима от них. При это между собой признаки независимы. При зависимости между признаками возникает проблема мультиколлинеарности, а за ней и проблема переобучения.

### Выбранные признаки.

Для расхода газа (Fг): давление воды после экономайзера (Pвэ), давление пара в барабане (Рпб), положение исполнительного механизма газа (ИМГ), частота вращения вентилятора дымососа (fдым), давление в дымоходе после экономайзера (Pдэ), общее давление газа (Рго).

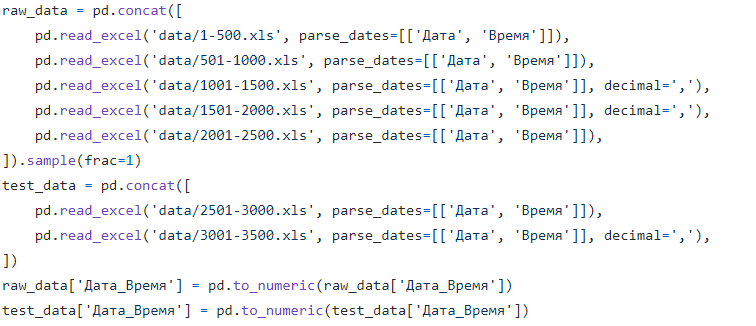
Для расхода воды (Fв): давление воды после экономайзера (Pвэ), давление пара в барабане (Рпб), температура воды после экономайзера (Tвэ), положение исполнительного механизма воды (ИМП), температура дымовых газов после котла (Tдк).

Для содержания оксида углерода (СО): температура газа (Tг), Разрежение в дымоходе после экономайзера (Pдэ), общее давление газа (Рго), частота вращения вентилятора дымососа (fдым), дата и время (Дата\_Время).

Для содержания кислорода после котла ('О2к'): температура газа (Tг), Разрежение в дымоходе после экономайзера (Pдэ), общее давление газа (Рго), частота вращения вентилятора дымососа (fдым), дата и время (Дата\_Время).

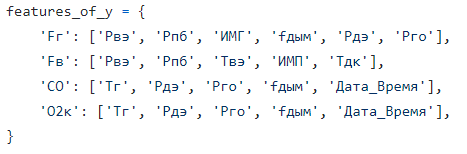
### Программная реализация.

В предыдущем подразделе было выявлено, что дата и время в признаках улучшает модели предсказания содержания оксида углерода и кислорода. А также то, что предсказание этих параметров нуждается в предварительном разбиение выборок и только потом перемешивания обучающей выборки. Поэтому дата и время добавлены в новых признаках. Для преобразования их к числовому виду применяется метод to\_numeric из библиотеки pandas. Также ранее дата и время были обозначены как колонка индексов для данных, теперь же следует убрать это обозначение при чтении данных, индексы будут присвоены автоматически, от 0 по возрастанию. Также на рис.8 представлено новое разбиение данных. Для обучающей выборки берется 2500 строк (первые 5 файлов), а для тестовой таким же способом 1000 строк (оставшиеся 2 файла). Обучающая выборка перемешивается методом *sample* с параметром *frac* равным единице.



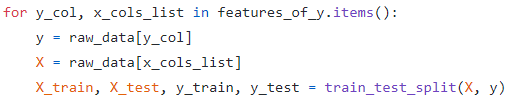
Программная реализация 2.1.

В связи с тем, что для каждой целевой переменной теперь выбраны свои признаки объявление массивов как в первом подразделе не подходит. Для нового подхода была применена структура данных языка *python* – словарь. В словаре ключом является целевая переменная, а значением массив обозначений признаков. Также объявление матрицы признаков теперь не может происходить сразу.



Программная реализация 2.2.

Далее также идет цикл, но уже не по элементам массива обозначений целевых признаков, а по парам ключ-значение словаря. Для каждого ключа создается матрица *y* а по значению матрица *X*. По этим матрицам создаются обучающие и тестовые выборки. Далее код не отличается от кода в первом подразделе.



Программная реализация 2.3.

### Новые показатели моделей

Для новых признаков были также обучены модели линейных регрессия и градиентного бустинга. Новые значения представлены в таблице 2.1.

Таблица 2.1 – Качество моделей с новыми признаками.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | | | |
| Целевая переменная | CO | O2к | Fв | Fг |
| Линейная регрессия | -0,37 | -0,87 | 0,81 | 0,78 |
| Линейная регрессия с L1 и L2 регуляризациями | 0,2 | 0 | 0,87 | 0,66 |
| Линейная регрессия с L1 и L2 регуляризациями и кросс-валидацией | -1,2 | 0 | 0,87 | 0,64 |
| Градиентный бустинг | 0,53 | -0,31 | 0,94 | 0,73 |

­

### Анализ

Основная цель этого подраздела – улучшение существующих моделей. По итогам прошлого подраздела модели предсказания кислорода и углекислого газа совсем не справлялись со своей задачей. Теперь видно, что модель градиентного бустинга по предсказанию углекислого газа может считаться приемлемой в какой-то степени, так как коэффициент детерминации 0,53. Но модели предсказания содержания кислорода все еще полностью не рабочие. При этом модель предсказания расхода воды методом градиентного бустинга стала лучше. Коэффициент детерминации увеличился с 0,91 на 0,94. Три сотые не сильное изменение, но при близости показателя к единице это показывает достаточно существенное влияние нового подхода. В свою очередь предсказание расхода газа улучшилось в среднем на 0,2. При этом линейная регрессия предсказывает даже лучше градиентного бустинга.

### Вывод

По итогу не получилось обучить хорошую модель по предсказанию содержания газов. Но предсказание расхода газа и воды улучшилось. Модель предсказания расхода газа не совсем соответствует критерию, при котором модель можно считать хорошей, если коэффициент детерминации больше или равен 0,8, но коэффициент близок к этому значению. При этом в дальнейшем будет считаться значение расхода газа, деленное на расход воды. А так, как модель предсказания расхода воды достаточно близка к идеальной, значение, которое будет использоваться можно считать приемлемым.

## Минимизация расхода газа

### Задача

При помощи обученных моделей предсказывать значения расхода газа и расхода воды для измененных признаков. Находить значения признаков, при которых отношение расхода газа к расходу воды минимальное, при этом давление в барабане неизменно.

### Анализ возможного решения

Возможный способ работы программы был в том, чтобы для каждой новой строки находить минимум, корректирую часть параметров. Такой способ, к сожалению, занимает огромное количество ресурсов и, в связи с этим выполняется слишком большое количество времени. Так как данные собираются каждую минуту, нужно предсказывать как изменить какие параметры, что бы в следующую минуту отношение расхода газа к расходу воды было минимальным.

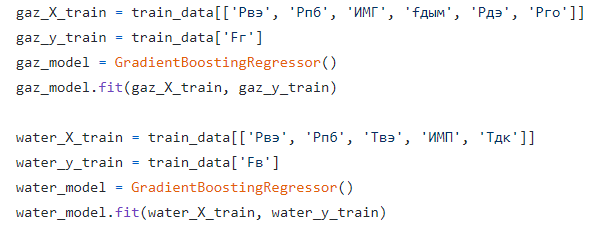
Первый способ уменьшить время выполнения – это проанализировать влияние параметров, чтобы найти зависимости.

Второй – проанализировав влияние параметров, уменьшить диапазон или шаг значений параметра, для проверки.

### Программная реализация

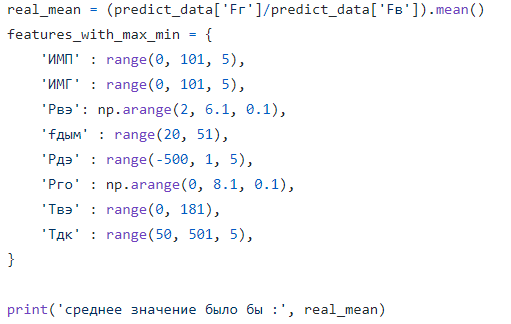
Разбиение данных остается таким же. Тестовая выборка состоит из 1000 строк. Но теперь тестовая выборка будет использоваться не для определения качества модели, а в качестве использования вместо новых данных, которые возможно могут приходить каждую минуту в дальнейшем.

Модели теперь только две. Одна для предсказания расхода газа и одна для предсказания расхода воды. Обе модели использую метод градиентного бустинга, т.к. этот метод показал себя как наиболее качественный и показывал лучший или приближенный к нему результат на каждом шаге. Это дает уверенность в модели при возможных изменениях в данных или подходе.



Программная реализация 3.1.

Сразу высчитывается и выводится среднее значение отношения расхода газа к расходу воды в тестовой выборке, для дальнейшего сравнения. Также заранее определяется словарь, в котором для каждого признака создается массив значений от минимума к максимуму. Возможные теоретические и реальные минимумы и максимумы значений представлены в таблице 1.1.



Программная реализация 3.2.

Для нахождения значения параметра для минимума отношения создан цикл по всем элементам словаря признаков *features\_with\_max\_min*. Для дальнейшего анализа вначале инициализируются переменные:

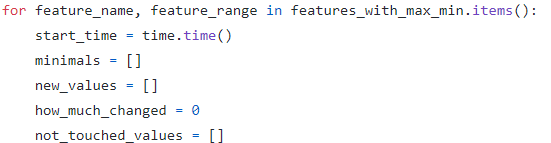
*start\_time* – для расчета потраченных на вычисление секунд,

*minimals* – массив всех значений минимумов отношения для каждой строки,

*new\_values* – массив, содержащий только значения измененных параметров, для каждой строки,

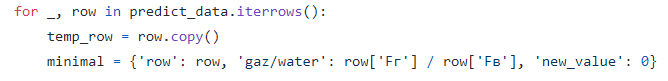
*how\_much\_changed* – переменная счетчик, для подсчета количества строк, которые были изменены,

*not\_touched\_values* – массив значений параметров в строках, в которых не был изменен параметр,



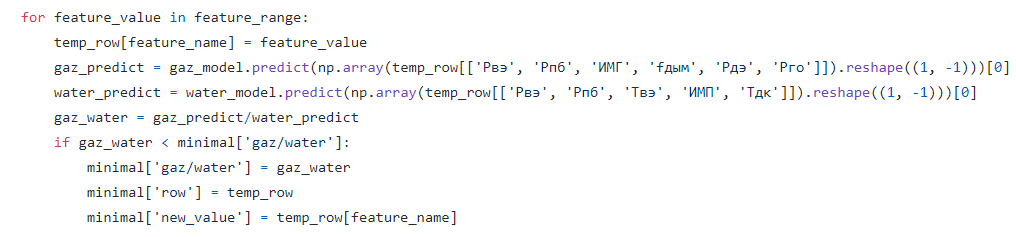
Программная реализация 3.3.

Дале цикл проходится по всем строкам тестовой выборки. Для каждой строки будет определяться отношение газа к воде и минимум отношения. Для этого создается переменная *temp\_row* временной строки, которая является копией строки на текущей итерации и переменная *minimal*, которая является словарем, который содержит значение изначальной строки, значение отношения и новое значение, если оно было присвоено.



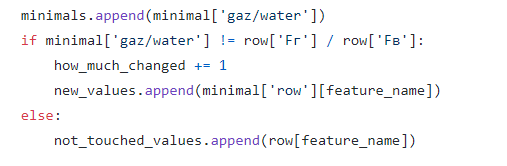
Программная реализация 3.4.

Для определения при каких значениях параметрах будет достигнут минимум отношения создан цикл по возможным значениям признака. В нем сначала меняется значение параметра в строке. Потом по параметрам, в том числе измененному, предсказываются значения расхода газа и расхода воды. Высчитывается отношение, и если отношение меньше записанного в словарь *minimals* минимума, то все значения словаря *minimals* обновляются.



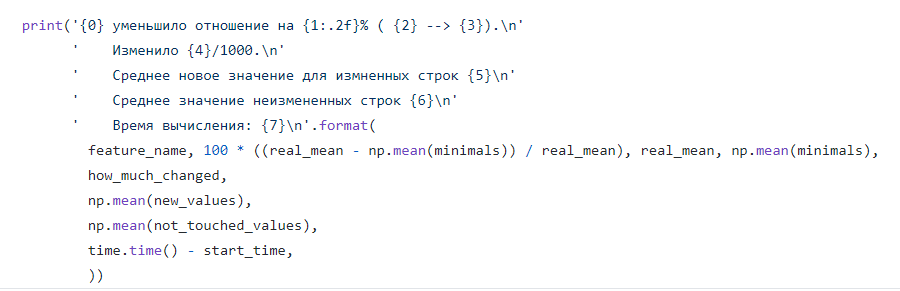
Программная реализация 3.4.

После чего заполняется массив минимумов и если значение минимума в словаре *minimals* было изменено, то добавляется счетчик изменений и значение в массив *new\_values*, иначе добавляет значение в массив *not\_touched\_values*.



Программная реализация 3.6.

После чего выводятся значения для дальнейшего анализа.



Программная реализация 3.7.

### Принципиальное представление

Сначала обучаются модели машинного обучения, на основе выводов первых двух подразделов. Далее, для каждого из параметров, для которых нужно провести анализ, выполняются следующие шаги:

* Представляется, что есть 1000 новых строк, т.к. для модели эти строки действительно новые.
* Для каждой строки выполняется исследование:
  + Параметр изменяется в рамках своих пределов.
  + Обученные модели предсказывают значения расхода газа и расхода воды.
  + Находится значение параметра, при котором отношение расхода газа к расходу воды было бы минимальным для текущей строки (при неизменных значениях остальных параметров).
* Для дальнейшего анализа рассчитывается:
  + В скольких строках из 1000 отношение могло бы быть меньше, если бы параметр принимал другое значение.
  + В тех случаях, когда значение отношения могло быть меньше, какое среднее среди новых, оптимальных, значение параметра. Для понимания, какое новое значение чаще всего является оптимальным.
  + В тех случаях, когда значение отношения Fг/Fв было оптимальным, какое среднее значение было у параметра.
  + Среди новых 1000 строк, с минимизированным значением отношения в каждой строке, какое среднее отношение расхода газа к расходу воды.
  + Насколько новое среднее отношение больше (в процентах) отношения, если бы изменения параметров не производились.

### Нахождение значений параметров для минимизации

Вышеуказанный код был выполнен для каждого параметра. Для каждого из параметров в таблице 2.4 записаны получившиеся средние значения отношения после минимизации. Что показывает то, насколько в среднем можно уменьшить значение отношения изменяя каждый параметр по отдельности.

Результаты выполнения вышеуказанного кода в табличном виде:

Таблица 3.1 Влияния признаков на отношение Fг/Fв.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Признак | Строк изменено | Среднее значение измененных параметров | Среднее значение не измененных параметров | Новое значение отношения Fг/Fв | Насколько меньше значения отношения без изменения |
| Без изменений | 0 | - | - | 6,79 | 0% |
| ИМП | 847 | 100 | 74,54 | 4,2 | 38,04% |
| Pдэ | 890 | 0 | -280,64 | 5,84 | 14,01% |
| ИМГ | 796 | 100 | 74,81 | 5,99 | 11,72% |
| Pго | 649 | 8 | 4,82 | 6,19 | 8,76% |
| Tвэ | 740 | 180 | 110,79 | 6,2 | 8,69% |
| Pвэ | 608 | 6 | 4,99 | 6,35 | 6,39% |
| Tдк | 590 | 500 | 129,09 | 6,36 | 6,26% |
| fдым | 540 | 50 | 25 | 6,45 | 4,94% |

Из таблицы видно, что во всех случаях минимальное значение отношения достигается путем максимизации признака.

Также из графика на рис.18 видно, что значение отношения увеличивается в точности тогда, когда значение положения исполнительного механизма воды (ИМП) уменьшается. При этом одна только видимая зависимость на графике не дает точного понимания, что зависимость существует. Благодаря машинному обучению можно определить эту зависимость с учетом других параметров, так как модель обучается на множестве признаков и учитывает их при предсказании значения.

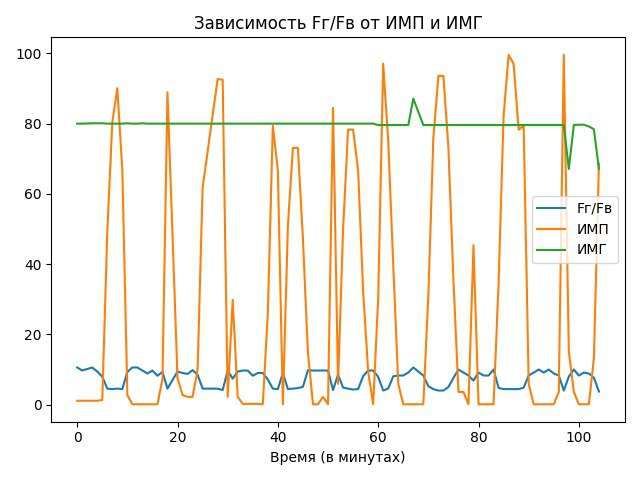


График отношения Fг/Fв, ИМП и ИМГ

### Анализ

Основным значением в таблице является процент изменения отношения расхода газа к расходу воды. Так как это решает напрямую задачу минимизации этого отношения. При это можно заметить, что параметр частоты вентилятора дыма в самом низу по этому параметру. Что не значит, что этот параметр не влияет на отношение Fг/Fв, возможно этот параметр уже достаточно оптимален для минимизации отношения и его изменение не так сильно влияет.

Самое важное, что все средние значения измененных параметров равны максимуму значения параметра. Что означает, что для минимизации отношения достаточно стараться максимизировать эти параметры.

Основными параметрами являются ИМП и ИМГ, так как они напрямую влияют на расход газа и расход воды, имеют большой потенциал для улучшения и легко могут изменяться системой, так как это значения положений механизмов, которые задаются на котельной. При этом из графика на рисунке 18 видно, что в основном значения этих параметров стремятся быть высокими. Значения ИМГ всегда в промежутке от 60 до 80, а значения ИМП либо приближены к высоким показателям ИМГ, либо близки к 0. И то и то может быть обусловлено технологическими ограничениями, фундаментальными или временными. Возможно потенциал улучшения этих параметров невозможно реализовать, тогда стоит обратить на другие параметры таблицы 3.1.

### Вывод

По результатам анализа найдены основные параметры, на которые стоит обратить внимание. При этом внедрение нового программного обеспечения не обязательно. Получившиеся значения могут быть изучены специалистами в предметной области и применены для оптимизации процесса сжигания газового топлива в парогенераторе.

Заключение

По итогу исследования изучены возможности оптимизации модели сжигания газового топлива в парогенераторе, влияние параметров на критерий оптимизации и степень возможного улучшения модели для каждого из изученных параметров. Что может успешно повысить эффективность парогенератора. Каждая из поставленных задач была выполнена и существенно повлияла на ход исследования. Кроме задачи учета количества выбросов при оптимизации, которая рассматривалась как возможность, но не была реализована, в связи с трудностью предсказания параметров выбросов и неудачей в построении моделей для этих параметров. Так были исследованы представленные данные, по ним построены первые модели, далее модели были проанализированы и улучшены и наконец использовались для нахождения принципов энергетической оптимизации котлоагрегата. Результаты рекомендуется сначала рассмотреть специалистами предметной области, для определения возможности их применения в реальных условиях. При этом разработанные программы внедрения не требуют. Внедрение возможно при необходимости, для этого потребуется небольшие доработки, обусловленных самим процессом внедрения и индивидуальными в каждом случае.

Успешно определено, что машинное обучение можно использовать не только для предсказания целевой переменной при текущих данных, но и для предсказания на множестве вариантов измененных данных, чтобы определить, что и как можно изменить для лучшего результата. Для больших и сложных данных это может потребовать намного больших ресурсов. Даже в данной работе каждый параметр анализировался отдельно, в связи с ограниченностью вычислительных ресурсов. Но при грамотном подходе и достаточно точной модели, такой подход может быть очень выгодным, благодаря предоставлению новой возможности анализирования систем.

Список литературы

1. Новиков, О. Н. Как повысить эффективность работы котельной / О.Н. Новиков, А.Н. Окатьев, Н.А. Антонова. // Промышленно–строительное обозрение – 2013 – № 8 – с.61–63.
2. Разинков Е.В. Машинное обучение // Электронная библиотека ИВМиИТ-КФУ. – 2015. – 25 апреля [Электронный ресурс]. URL: <https://vmkhelp.ru/wp-content/uploads/2017/09/machine_learning_at_kfu_25_04_15.pdf> (дата обращения: 7.05.2020).
3. Воронцов К.В. Лекции по алгоритмическим композициям // Профессиональный информационно-аналитический ресурс. – 2012. – 7 октября [Электронный ресурс]. URL: <http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf> (дата обращения: 7.05.2020).
4. Кашницкий Ю. С., Игнатов Д. И. Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов // Интеллектуальные системы. Теория и приложения. – 2015. – Т. 19. – №4. – С.45 – 47.