Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт
Высшая школа теоретической механики и математической физики

Работа допущена к защите
Директор ВШТМиМФ
д.ф.-м.н., чл.-корр. РАН
_____ А.М. Кривцов
«_____» _____ 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

ПРОГНОЗИРОВАНИЕ КРИТИЧЕСКОЙ ТЕМПЕРАТУРЫ СВЕРХПРОВОДНИКОВ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

по направлению 01.03.03 Механика и математическое моделирование
по образовательной программе
01.03.03_03 Математическое моделирование процессов нефтегазодобычи

Выполнил
студент гр. 5030103/00301 А.В. Корнелюк

Руководитель
Доцент ВШТМиМФ, к.т.н. О.А. Троицкая

Консультант
Ассистент ВШТМиМФ А.Д. Ершов

Санкт-Петербург
2024

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт

Высшая школа теоретической механики и математической физики

УТВЕРЖДАЮ

Директор ВШТМиМФ

А. М. Кривцов

«__»_____20__г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студенту Корнелюк Алексею Викторовичу, гр. 5030103/00301

1. Тема работы: Прогнозирование критической температуры сверхпроводников с применением методов машинного обучения

2. Срок сдачи студентом законченной работы: 30.05.2024

3. Исходные данные по работе: актуальные научные публикации по теме работы, набор данных о физических и химических свойствах сверхпроводников и их критических температурах, набор данных о химических формулах сверхпроводников.

4. Содержание работы (перечень подлежащих разработке вопросов): анализ существующих в настоящее время методов вычисления критической температуры сверхпроводников. Составление математической модели для вычисления критической температуры с использованием различных методов машинного обучения. Оценка эффективности работы этих методов и сравнение результатов их работы с реальными данными об известных критических температурах сверхпроводников.

5. Перечень графического материала (с указанием обязательных чертежей): не предусмотрено

6. Консультанты по работе: Ершов А.Д. – ассистент ВШТМиМФ

7. Дата выдачи задания 26.02.2024

Руководитель ВКР _____ Троицкая О.А. – доцент ВШТМиМФ, к.т.н., доцент

Задание принял к исполнению 26.02.2024

Студент _____ Корнелюк А.В.

РЕФЕРАТ

На 44 с., 11 рисунков, 6 таблиц, 0 приложений.

КЛЮЧЕВЫЕ СЛОВА: СВЕРХПРОВОДИМОСТЬ, СВЕРХПРОВОДНИК, МАШИННОЕ ОБУЧЕНИЕ, СТАТИСТИЧЕСКОЕ ОБУЧЕНИЕ, КРИТИЧЕСКАЯ ТЕМПЕРАТУРА.

Тема выпускной квалификационной работы: «Прогнозирование критической температуры сверхпроводников с применением методов машинного обучения»

В данной работе изложена сущность подхода к созданию статистической модели для предсказания критической температуры сверхпроводника на основе характеристик, извлеченных из химической формулы сверхпроводника. Статистическая модель, созданная с помощью метода машинного обучения, а именно градиентного бустинга, дает разумные по точности предсказания критической температуры: ± 8.89 К по среднеквадратичной погрешности (RMSE). Характеристики, извлеченные на основе теплопроводности, атомного радиуса, атомной массы, валентности и сродства к электрону материала сверхпроводника вносят наибольший вклад в точность предсказания модели. Важно отметить, что модель не предсказывает, является ли материал сверхпроводником или нет; она только дает предсказания критической температуры для сверхпроводников. Результаты работы могут быть полезны исследователям, занимающимся развитием теории сверхпроводимости. Например, модель помогает выделить свойства материала, которые наиболее сильно влияют на его критическую температуру. Возможно развитие работы путем добавления большего количества признаков о материалах в обучающие данные, либо путем применения более сложных моделей машинного обучения, таких как нейронные сети.

ABSTRACT

44 pages, 11 figures, 6 tables, 0 appendices

KEYWORDS: SUPERCONDUCTIVITY, SUPERCONDUCTOR, MACHINE LEARNING, STATISTICAL LEARNING, CRITICAL TEMPERATURE.

The subject of the graduate qualification work is «Predicting the critical temperature of superconductors using machine learning methods».

This paper outlines the essence of an approach to create a statistical model for predicting the critical temperature of a superconductor based on features extracted from the chemical formula of the superconductor. The statistical model created using a machine learning method, namely gradient bousting, yields reasonable accuracy in predicting the critical temperature: ± 8.89 K in terms of root mean square error (RMSE). Characteristics extracted from the thermal conductivity, atomic radius, atomic mass, valence, and electron affinity of the superconductor material contribute most to the accuracy of the model prediction. It is important to note that the model does not predict whether a material is a superconductor or not; it only provides critical temperature predictions for superconductors. The results of the work may be useful to researchers involved in developing the theory of superconductivity. For example, the model helps to highlight the properties of a material that most strongly influence its critical temperature. It is possible to develop the work by adding more materials features to the training dataset, or by applying more complex machine learning models such as neural networks.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Сверхпроводящие материалы, также известные как сверхпроводники - это материалы, которые проводят электрический ток с нулевым сопротивлением. Сверхпроводники имеют широкое и значительное практическое применение. Одно из наиболее известных таких применений - в системах магнитно-резонансной томографии (МРТ), которые широко используются медицинскими работниками для получения детальных снимков внутренних органов. Другие известные области применения включают сверхпроводящие катушки, используемые в Большом адронном коллайдере для поддержания высоких магнитных полей, где было подтверждено существование бозона Хиггса, а также чрезвычайно чувствительные устройства для измерения магнитного поля, называемые устройствами сверхпроводящей квантовой интерференции (SQUID). Кроме того, сверхпроводники имеют потенциал изменить будущее энергетической отрасли, поскольку сверхпроводящие провода и электрические системы с нулевым сопротивлением могут передавать и поставлять электрический ток без потерь энергии [6].

Однако широкому применению сверхпроводников препятствуют две основные проблемы: (1) Сверхпроводник проводит ток с нулевым сопротивлением только при критической температуре сверхпроводимости (T_c) или ниже нее. Критическая температура сверхпроводимости различается для различных материалов. Выполнение этого требования весьма непрактично, так как сверхпроводник должен быть охлажден до чрезвычайно низких температур, близких к температуре кипения азота (77 К) или ниже нее, прежде чем он приобретет свойство нулевого сопротивления [1]. (2) Научная модель и теория, предсказывающие T_c, остаются открытыми проблемами, которые ставят в тупик научное сообщество с самого открытия сверхпроводимости в 1911 году Хайке Камерлингом Оннесом.

В отсутствие каких-либо теоретических моделей прогнозирования критической температуры сверхпроводимости простые эмпирические правила, основанные на результатах экспериментов, на протяжении многих лет помогали исследователям в синтезе сверхпроводящих материалов. Например, выдающийся физик-экспериментатор Маттиас [8] пришел к выводу, что T_c зависит от количества доступных валентных электронов на атом (некоторые из полученных им правил стали известны как правила Маттиаса). Тем не менее, в настоящее время известно, что многие из простых эмпирических правил нарушаются [4].

В данной работе используется подход, полностью основанный на данных, для создания статистической модели, которая предсказывает T_c сверхпроводника на основе его химических и физических свойств. Данные о сверхпроводниках получены из Базы данных по сверхпроводящим материалам, поддерживаемой Японским национальным институтом материаловедения (Japan's National Institute for Materials Science, NIMS). Данные находятся в сети Интернет [19].

Существует несколько аналогичных работ, в которых основное внимание уделяется статистическим моделям для прогнозирования критической температуры сверхпроводимости T_C для широкого класса материалов [7; 9; 10]. Однако, авторы двух из упомянутых работ [9; 10] сосредоточили свои усилия на прогнозировании T_c только для сверхпроводников на основе Fe и MgB2 соответственно. В данной же работе ставится цель прогнозировать T_c для широкого класса материалов, не ограничиваясь двумя подклассами, упомянутыми выше. От подхода автора третьей работы [7] подход данной работы отличается следующим образом: в данной работе используется больший набор данных; также используется одна большая модель для получения прогнозов, а не каскад небольших менее сложных моделей; создается большее количество признаков только на основе элементарных химических и физических свойств сверхпроводников.

Целью данной работы является создание модели машинного обучения, позволяющей количественно спрогнозировать значение критической температуры сверхпроводимости для материала, основываясь на его химических и физических свойствах.

Результатом работы модели будет являться оценка критической температуры для конкретного сверхпроводника, которую можно в дальнейшем использовать для проведения эксперимента эмпирического измерения критической температуры данного сверхпроводника. Предполагается, что данная модель и предоставляемая ею оценка будет носить рекомендательный характер при определении диапазона температур для дальнейшей проверки точного значения критической температуры сверхпроводимости, и учет ее результатов позволит сократить временные ресурсы специалистов, занимающихся работой с новыми сверхпроводниками [2]. К тому же, экспериментальное определение T_c может быть дорогостоящей или трудоемкой задачей, если не сужать границы поиска T_c до определенных пределов, в которых затем и проводить эксперимент.

Таким образом, можно выделить следующие основные задачи, которые решаются в данной работе:

- Подготовка данных и работа с признаками - химическими и физическими свойствами сверхпроводников
- Определение подхода для реализации модели прогнозирования критической температуры
- Построение различных моделей машинного обучения для решения задачи
- Проведение анализа результатов различных моделей и сопоставление эффективности моделей для выбора наилучшей в решении данной задачи

ГЛАВА 1. ПОДГОТОВКА ДАННЫХ И РАБОТА С ПРИЗНАКАМИ

В данной главе описаны шаги по подготовке данных и выделению признаков. В параграфе 1.1 описывается получение и обработка данных по химическим элементам, которые входят в состав материалов сверхпроводников. На основе этих данным будет сформирована часть признаков. В параграфе 1.2 описывается подготовка данных, взятых из Базы знаний сверхпроводящих материалов, поддерживаемой Японским национальным институтом материаловедения (Japan's National Institute for Materials Science, NIMS). В параграфе 1.3 описывается процесс создания признаков из данных, полученных согласно параграфам 1.1 и 1.2.

1.1. Подготовка данных о химических элементах

Данные о химических элементах с 46 признаками и 86 строками (соответствующими 86 элементам) получены с помощью функции Element Data из Wolfram and Research от Mathematica [15]. Около 12% из 46 x 86 = 3956 записей отсутствуют.

При выборе свойств элементов учитываются рекомендации из источника [4]. Например, из данных исключена переменная «температура кипения» и вместо нее используется переменную «теплота плавления», которая в отличие от «температуры кипения» не имеет пропущенных значений, при этом она сильно коррелирует с переменной «температура кипения», так что исключение обосновано.

По итогам отбора признаков элементов получен список из 8 свойств, показанных в табл.1.1.

Таблица 1.1

Свойства элементов, используемые для создания признаков для прогнозирования T_c

Свойство	**Ед. измерения**	**Описание**
Атомная масса	а.е.м.	Суммарная масса покоя протонов и нейтронов
Первая энергия ионизации	кДж/моль	Энергия для отрыва валентного электрона от атома
Атомный радиус	пм (10^{-12} м)	Радиус атома
Плотность	кг/м^3	Плотность при нормальных условиях
Сродство к электрону	кДж/моль	Энергия для присоединения электрона к нейтральному атому
Теплоемкость плавления	кДж/моль	Энергия для перехода из твердого состояния в жидкое
Теплопроводность	Вт/(м · К)	Коэффициент теплопроводности
Валентность	-	Число химических связей, образуемых элементом

Далее необходимо подготовить данные о сверхпроводниках, чтобы затем объединить их с полученными данными о химических элементах.

1.2. Подготовка данных о сверхпроводниках

База данных по сверхпроводящим материалам поддерживается NIMS, государственным учреждением, базирующимся в Японии. База данных содержит большой список сверхпроводников, их критические температуры и ссылки на источники, в основном из журнальных статей [19]. Насколько известно в рамках поисков базы данных для данной работы, это самый полный источник информации по сверхпроводникам. Для доступа к базе данных требуется логин и пароль, но это обеспечивается простым процессом регистрации. Зарегистрированные пользователи могут просматривать и загружать базу данных.

После входа в систему был выбран раздел «OXIDE & METALLIC». На рис.1.1 показан внешний вид интерфейса поиска. Чтобы получить все данные, необходимо нажать на кнопку «поиск». Из базы данных было получено 31 611 строк в формате файла, разделенного запятыми (Comma Separated File, CSV). Ключевые необходимые столбцы для данной работы в этом файле следующие: «элемент», «химическая формула материала», и «Tc» - критическая температура сверхпроводимости для данного материала. Переменная «num» является уникальным идентификатором для каждой строки. Столбец «refno» содержит ссылки на источники, из которых получены данные о свойствах материала.

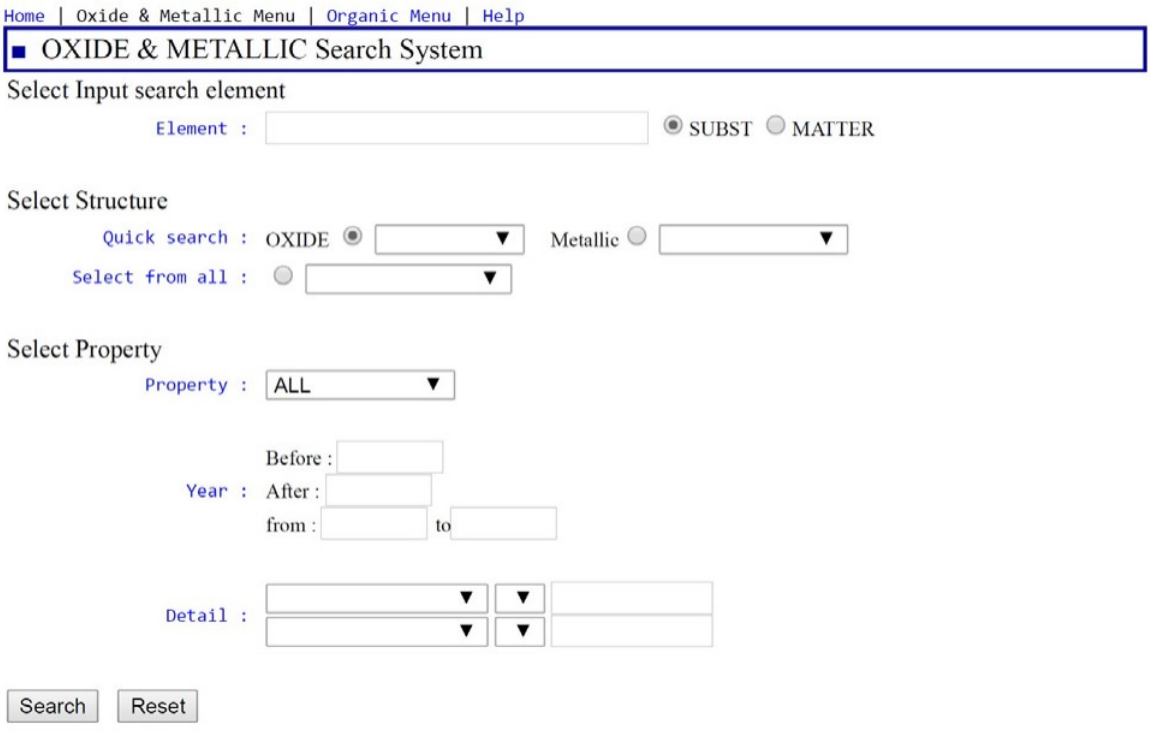

Рис.1.1. Внешний вид интерфейса поиска базы данных NIMS

Следующие шаги описывают процесс предварительной подготовки и очистки данных. Это важная задача, так как именно от качества имеющихся в распоряжении исследователя данных напрямую зависит качество прогнозов модели машинного обучения.

Итак, процесс подготовки и очистки включает в себя последовательные действия:

A. Удаление всех столбцов от «ma1» до «mj2» включительно;
B. Сортировка данных по убыванию T_c (для удобства);
C. Значения критической температуры для переменных со следующими значениями колонки «num» ошибочно сдвинуты на один столбец вправо: 31 020, 31 021, 31,022, 31 023, 31 024, 31 025, 153 150, 153 149, 42 170, 42 171, 30 716, 30 717, 30 718, 30 719, 150 001, 150 002, 150 003, 150 004, 150 005, 150 006, 150 007, 30 712, 30 713, 30 714, 30 715. Эти ошибки были исправлены сдвиганием значений в столбце для этих строк;
D. Удаление данных с ошибочно записанной критической температурой; они содержат значения T_c выше 203 К, что по состоянию на дату проведения исследования было самой высокой достоверно зарегистрированной критической температурой. La0.23Th0.77Pb3 (num = 111 620), Pb2C1Ag2O6 (num = 9 632), Er1Ba2Cu3O7-X (num = 140);
E. Удаление всех строк, в которых значение столбца «Tc» = 0 или отсутствует;
F. Удаление столбцов «nums», «mo1», «mo2», «oz», «str3», «tcn», «tcfig», «refno»;
G. Ручное изменение всех материалов с формулой содержания кислорода, такой как O7-X, на наилучшее приближение к содержанию кислорода. Например, O7-X заменяется на O7, O5+X заменяется на O5 и т.д. Это, безусловно, вносит некоторую погрешность в данные, но невозможно просмотреть документ за документом из источников данных для всех строк в базе данных, чтобы получить более точные оценки содержания кислорода.
H. Использование статистического программного обеспечения на языке программирования и вычислительного пакета CHNOSZ [5] для проверки достоверности химических формул. В пакете имеется функция `makeup`, которая считывает химическую формулу в строковом формате и разбивает формулу на элементы и их соотношения. Она иногда выдает

ошибку или предупреждение, когда химическая формула не имеет смысла. Например, при проверке формулы $Pb_{-2}O$ выдается предупреждающее сообщение: Отрицательное значение Pb не имеет смысла. Однако функция не проверяет, может ли материал существовать на самом деле. С помощью пакета CHNOSZ вносятся следующие изменения:

1. Удалены сверхпроводники с формулами Yo975Yb0.025Ba2Cu3O, Yo975Yb0.025Ba2Cu3O, Yo975Yb0.025Ba2Cu3O. Не существует химического элемента с символом Yo. Вероятно, что Y0.975 был ошибочно записан как Yo975, но нельзя утверждать однозначно;
2. Удалены сверхпроводники с формулами Bi1.7Pb0.3Sr2Ca1Cu2O0, La1.85Nd0Ca1.15Cu2O5.99, Bi0Mo0.33Cu2.67Sr2Y1O7.41, Y0.5Yb0.5Ba2Sr0Cu3O7, так как некоторые элементы имеют нулевые коэффициенты;
3. Удален сверхпроводник с формулой Y2C2Br0.5!1.5. Восклицательный знак провоцирует сообщение об ошибке и не может быть однозначно интерпретирован;
4. Удален сверхпроводник с формулой Y1Ba2Cu3O6050. Коэффициент 6050 для кислорода, вероятно, является ошибкой;
5. Удален сверхпроводник с формулой Hg1234O10. Коэффициент 1234 для ртути, вероятно, является ошибкой;
6. Удалены сверхпроводники с формулами Nd185Ce0.15Cu1O4 удален. Коэффициент 185 для неодима, вероятно, является ошибкой. В данных уже содержится Nd1.85Ce0.15Cu1O4;
7. Изменено значение формулы Bi1.6Pb0.4Sr2Cu3Ca2O1013 на Bi1.6Pb0.4Sr2Cu3Ca2O10.13, поскольку соседние с ней строки данных содержат формулы с значением O10.xx;
8. Изменено значение формулы Y1Ba2Cu285Ni0.15O7 на Y1Ba2Cu2.85Ni0.15O7, поскольку соседние с ней строки в данных содержат формулы с Cu2.xx.

I. Названия столбцов «Tc» и «element» изменены на «critical_temp» и «material» соответственно. Это сделано для удобства дальнейшей работы и никак не влияет на содержание данных.

По итогам очистки данных исключено 6 750 строк. Таким образом, остается 24 861 строк данных. Остальная часть подготовки данных выполняется с помощью языка программирования R. Исключаются все сверхпроводники, в состав которых

входит элемент с атомным номером больше 86. Это позволяет исключить дополнительные 973 строки данных. Например, исключены сверхпроводники, содержащие уран.

Также удаляются повторяющиеся строки, так как в данной работе невозможно вручную проверить, являются ли повторяющиеся строки подлинными независимыми результатами из независимых экспериментов или это просто повторяющиеся строки, являющиеся ошибкой. В конце концов после подготовки и очистки данных получено 21 263 строки данных, что составляет около 67% от исходных данных.

Теперь все данные собраны, и следующий шаг - подготовка признаков для модели на основе этих данных.

1.3. Подготовка признаков на основе обработанных данных

В этом разделе описан процесс выделения признаков на примере материала из двух химических элементов: рассматривается материал Re_6Zr_1 при T_c = 6,7 К. Пример иллюстрирует расчет характеристик, выделенных на основе теплопроводности.

Коэффициенты теплопроводности рения и циркония равны $t_1 = 48$ и $t_2 = 23$ - соответственно. Соотношение элементов в материале используется для определения характеристик:

$$p_1 = \frac{6}{6+1} = \frac{6}{7},\ p_2 = \frac{1}{6+1} = \frac{1}{7} \tag{1.1}$$

Аналогично получается соотношение коэффициентов теплопроводности в материале:

$$w_1 = \frac{48}{48+23} = \frac{48}{71},\ w_2 = \frac{23}{48+23} = \frac{23}{71} \tag{1.2}$$

Таким образом получатся пара значений, основанная на уравнениях (1.1) и (1.2):

$$A = \frac{p_1 \cdot w_1}{p_1 \cdot w_1 + p_2 \cdot w_2},\ B = \frac{p_2 \cdot w_2}{p_1 \cdot w_1 + p_2 \cdot w_2} \tag{1.3}$$

После получения значений p_1, p_2, w_1, w_2, A, B становится возможно выделить 10 характеристик теплопроводности рения и циркония, как показано в табл.1.2.

Далее описанный выше процесс повторяется с 8 переменными, перечисленными в табл.1.1. Например, для элементов, основанных на атомной массе,

коэффициенты теплопроводности t1 и t2 заменяются на атомные массы рения и циркония соответственно, затем по аналогии вычисляются p_1, p_2, w_1, w_2, A, B и, наконец, рассчитываются 10 признаков, определенных в табл.1.2. Таким образом, после выполненной подготовки признаков получается 8 x 10 = 80 признаков для каждого сверхпроводника. Еще один дополнительный признак - числовая переменная, подсчитывающая количество химических элементов в сверхпроводнике. В итоге получается в общей сложности 81 признак.

Таблица 1.2

Признаки элементов, полученные из свойств, для прогнозирования T_c на примере Re_6Zr_1

Признак	**Формула**	**Пример значения**
Среднее арифметическое	$= \mu = (t_1 + t_2)/2$	35.5
Взвешенное среднее	$= \nu = p_1 t_1 + p_2 t_2$	44.43
Среднее геометрическое	$= (t_1 t_2)^{1/2}$	33.23
Взвешенное среднее геометрическое	$= (t_1)^{p_1} (t_2)^{p_2}$	43.21
Энтропия	$= -w_1 ln(w_1) - w_2 ln(w_2)$	0.63
Взвешенная энтропия	$= -A ln(A) - B ln(B)$	0.26
Разность	$t_1 - t_2 \ \ (t_1 > t_2)$	25
Взвешенная разность	$p_1 t_1 - p_2 t_2$	37.86
Стандартное отклонение	$[(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]^{1/2}$	12.5
Взвешенное стандартное отклонение	$[p_1 (t_1 - \mu)^2 + p_2 (t_2 - \mu)^2]^{1/2}$	8.75

Итого, получены подготовленные и обработанные данные с 21 263 строками и 82 столбцами: 81 столбец соответствует извлеченным признакам и 1 столбец - наблюдаемым значениям T_c. Также рассмотрен, но не реализован подход, создающий функции, которые указывают, присутствует ли тот или иной элемент в сверхпроводнике или нет. Например, столбец, в котором указывалось бы, содержится ли, например, кислород в материале или нет. Этот подход отклонен из реализации по следующей причине: он добавляет к данным большое количество индикаторных переменных, из-за этого выбор и оценка модели становятся слишком сложными и при этом увеличивается вероятность переобучения под данные.

Таким образом, в этой главе подробно описан процесс сбора и подготовки данных, которые будут применяться для создания модели. Следующая глава посвящена определению машинного обучения, преимуществам моделей машинного обучения, а также критериям оценки эффективности этих моделей.

ГЛАВА 2. ОПРЕДЕЛЕНИЕ ПОДХОДА ДЛЯ РЕАЛИЗАЦИИ МОДЕЛИ

Глава посвящена выбору подхода к построению модели, введению в машинное обучение, различным архитектурам моделей машинного обучения, а также определению типа задачи и критериев оценки успешности решения данной задачи.

В параграфе 2.1 рассматриваются предпосылки к использованию методов машинного обучения для решения данной задачи и их преимущества. Параграф 2.2 определяет тип задачи в терминах методов машинного обучения. В параграфе 2.3 формулируется постановка задачи и критерии оценки качества решения.

2.1. Предпосылки к использованию методов машинного обучения

В настоящее время происходит значительный рост цифровизации, а также оптимизации бизнес-процессов и производственных мероприятий с помощью современных технологий и новых алгоритмов. Повсеместный переход в людей и компаний в цифровую среду способствует накоплению огромных объемов данных, особенно в науке и исследовательской деятельности. Один из наиболее эффективных в настоящее время методов анализа и обработки больших данных, а также автоматизации многих операций, заключается в применении машинного обучения для решения производственных задач.

Машинное обучение представляет собой раздел искусственного интеллекта, который фокусируется на разработке методов решения задач через обучение на примерах. Под примерами как правило подразумеваются определенные данные, в идеале - упорядоченные, например, в табличном виде. Основой всех методов машинного обучения служат элементы математической статистики, численных методов, методов оптимизации, теории вероятностей и математического анализа. Алгоритмы машинного обучения тренируются на предоставленных им данных, извлекая из них ключевые особенности и выявляя скрытые корреляции, которые невозможно обнаружить вручную человеческим взглядом из-за огромного объема данных или банального человеческого фактора. Этот навык нахождения закономерностей позволяет машинному обучению эффективно анализировать данные, запоминать полученную информацию, а затем строить прогнозы и выбирать наиболее подходящие решения из предложенных вариантов.

Машинное обучение принято разделять на два основных типа [20]:

- Обучение с учителем. Этот подход предполагает обучении моделей на заранее размеченных данных, где каждому примеру из данных соответствует метка (или значение), представляющая правильный ответ. Примеры задач обучения с учителем - классификация и регрессия.
- Обучение без учителя. При этом подходе модели обучаются на данных без заранее известных меток. Задачи обучения без учителя включают кластеризацию, снижение размерности и обнаружение аномалий. В задачах обучения без учителя перед моделью ставиться задача наиболее эффективно обрабатывать данные, не зная правильных ответов.

Отдельно стоит более подробно рассмотреть широко применяющиеся задачи регрессии и классификации, относящиеся к типу обучения с учителем:

- Регрессия – тип задачи, в которой обучающая выборка x_i характеризуется набором признаков, каждый из которых принимает вещественные значения. Прогноз модели y_i также принадлежит множеству вещественных чисел. Иными словами, в регрессии целевые данные являются непрерывно распределенными. При этом множество значений y_i не обязано быть конечным. Например, $y_i \in \mathbb{R}$.
- Классификация – тип задачи, при котором множество объектов принадлежит некоторому количеству классов, и распределяется по классам на основе некоторого характерного свойства или набора свойств. Выполнение классификации объекта означает указание номера или наименования класса, к которому он относится на основе его признаков. При этом множество возможных значений y_i является конечным и равно количеству заданных классов.

Таким образом, в настоящее время машинное обучение позволяет решать широкий спектр задач, улучшая качество анализа данных и автоматизируя многие процессы. Это способствует повышению эффективности и производительности работы в различных областях человеческой деятельности, и особенно помогает в исследованиях и научных работах.

В следующем параграфе определяется подходящий для данной работы тип задачи машинного обучения.

2.2. Определение типа задачи в терминах методов машинного обучения

Суть данной работы заключается в прогнозировании наиболее близкого к реальному значения критической температуры сверхпроводимости для определенных химических и физических свойств материала сверхпроводника на основе знания модели о критических температурах других сверхпроводников, также имеющих определенный набор параметров. В таком случае примером обучающих данных для модели становится информация о сверхпроводнике, а в качестве результата понимается непрерывно распределенная величина – значение критической температуры. Таким образом, исходя из постановки задачи, в данной работе для реализации модели выбран тип обучения с учителем, так как имеются обучающие размеченные данные, где известен правильный ответ.

Для определения типа задачи, которая решается в данной работе, проводится анализ данных, на основании которого делается выбор в пользу одного из типов. В ходе работы требуется определить критическую температуру сверхпроводимости, при этом критическая температура может принимать различные и не обязательно дискретно распределенные значения. Таким образом, количество возможных значений T_c не ограничено. На основании этого условия в качестве типа задачи принимается регрессия.

Наглядной иллюстрацией задачи регрессии является рис.2.1. На нем показана зависимость стоимости недвижимости от ее площади. Эта зависимость приближается линейной регрессией.

При обучении с учителем задается два множества: X – множество описаний объектов, Y – множество допустимых значений, и существует неизвестная целевая зависимость f, являющаяся отображением множества X на Y, иначе говоря:

$$X, Y \quad \exists f,\ f : X \to Y \tag{2.1}$$

Значения целевой функции $y_i = f(x_i)$ известны только на конечном числе объектов. Как правило, такая выборка объектов $x_i \in X, i = 1...m$ называется обучающей выборкой. Основной задачей является построение алгоритма α: $\alpha : \ X \to Y$, который приближал бы неизвестную целевую зависимость как на элементах обучающей выборки, так и на всем множестве возможных объектов класса X. Для измерения точности прогнозов модели определяют некоторый заданный функционал качества.

Рис.2.1. Пример задачи регрессии [16]

Таким образом, в данной работе решается задача регрессии. В следующем параграфе данная задача определяется с математической точки зрения.

2.3. Постановка задачи и определение критериев оценки качества решения

Как уже было отмечено ранее, в ходе данной работы решается задача регрессии. Данный алгоритм вычисляет ожидаемое значение целевой переменной на основе входных данных путем перемножения признаков на некоторые веса (т.е. коэффициенты), а затем определяет, насколько предсказанное значение близко к фактическому значению целевой переменной. После этого происходит переоценка весов в сторону более оптимального предсказания. Процесс повторяется итеративно до тех пор, пока не будет достигнут минимум функции, измеряющей ошибку алгоритма.

Рассматривается обучающая выборка $X^m = (x_1, y_1), \ldots, (x_m, y_m)$, где каждый набор данных содержит вещественнозначные значения: $(x_i, y_i) \in X^m \times Y$, $i = 1 \ldots m$, и где каждому набору признаков в виде вектора X, отвечающему за информацию о химических и физических свойствах материала, соответствует значение Y, являющееся вещественным числом, отвечающим за критическую температуру сверхпроводника.

В данной постановке предлагается оценивать значение целевой переменной на основе входных признаков. Полагается, что все объекты независимы и взяты из некоторого неизвестного распределения $(x_i, y_i) \in P(x, y)$, $i = 1...m$. Иными словами, информация о свойствах каждого из сверхпроводников никак не зависит от информации о других сверхпроводниках.

Поэтому требуется определить, во-первых, ожидаемое значение целевой переменной, а во-вторых, сделать это наиболее точно. Для оценки качества модели вводится понятие функционала потерь.

Функционал потерь указывает, насколько близко прогнозируемое значение к соответствующему ему истинному значению, или иначе говоря, насколько качественный с точки зрения отражения реальности получен результат. Чем больше прогнозируемое значение отклоняется от фактического значения, тем большее значение имеет функционал потерь. Таким образом, задача сводится к тому, чтобы с помощью алгоритма найти такое распределение исходов, чтобы предсказания соответствующих исходов были максимально точны, а выбранный функционал потерь – минимален. Это достигается за счет корректировки весов или некоторых правил, на основе которых модель прогнозирует результат.

В качестве функционала потерь в работе выбран корень из среднеквадратичной ошибки (Root Mean Squared Error, сокращенно RMSE), рассчитываемый по формуле:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{2.2}$$

где y_i - истинные значения, $\hat{y}_i$ - прогнозные значения модели, а N - число объектов в выборке, на которой обучается модель.

Величина RMSE отражает среднее расхождение между фактическими значениями и предсказанными значениями целевой переменной. Чем меньше значение RMSE, тем лучше качество модели. При этом разность между истинным и прогнозируемым значением возводится в квадрат, чтобы одинаково учитывать ошибки модели в обе стороны от истинного результата, а корень из суммы квадратов извлекается затем, чтобы иметь возможность интерпретировать качество результата в тех же единицах измерения, в которых измеряется целевая переменная Y. То есть если единица измерения температуры в данной задаче - Кельвин, то и величина RMSE будет иметь размерность Кельвин, и будет показывать, насколько в среднем модель ошибается в прогнозировании значения T_c.

Помимо RMSE, существуют также и другие метрики для оценки качества моделей регрессии. Рассмотрим некоторые из них, приведем их формулы и поясним преимущества и недостатки по сравнению с RMSE.

Среднеквадратичная ошибка (Mean Squared Error, MSE) вычисляется как среднее арифметическое квадратов ошибок предсказания:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Преимущества:

– MSE более чувствительна к большим ошибкам, так как возводит их в квадрат, что может быть полезно в задачах, где важны большие отклонения.

Недостатки:

– Как и RMSE, MSE достаточно чувствительна к выбросам, что может искажать оценку модели.
– MSE имеет те же единицы измерения, что и квадрат целевой переменной (а не сама целевая переменная), что делает интерпретацию сложнее.

Средняя абсолютная ошибка (Mean Absolute Error, MAE) вычисляется как среднее арифметическое абсолютных значений ошибок предсказания:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Преимущества:

– MAE также как и RMSE легко интерпретировать, так как она представляет собой среднюю ошибку в тех же единицах измерения, что и целевая переменная.
– MAE менее чувствительна к выбросам по сравнению с RMSE, так как большие значения, возведенные в квадрат, становятся еще больше. Благодаря этому свойству MAE лучше работает с задачах, где есть много выбросов или экстремальных значений в данных.

Недостатки:

– Как было сказано выше, MAE не возводит ошибки в квадрат, поэтому не выделяет большие ошибки так сильно, как RMSE. Поэтому использование MAE может создать модель, плохо реагирующую на появление сильных отклонений от предсказаний.

Коэффициент детерминации (R-squared, R^2) измеряет долю дисперсии зависимой переменной, которая объясняется моделью:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

где $\bar{y}$ – среднее значение целевой переменной.

Преимущества:

– R^2 предоставляет относительную меру качества модели, показывая, насколько хорошо модель объясняет данные по сравнению с предсказанием простого среднего значения по выборке.

Недостатки:

– R^2 может быть неинформативным для нелинейных моделей регрессии, где признаки входят в модель в степенях выше первой.
– Высокое значение R^2 не всегда означает хорошую модель, особенно при наличии переобучения, то есть подстраивания под обучающую выборку.

Таким образом, выбор метрики для оценки качества регрессии зависит от конкретной задачи и свойств данных. RMSE полезна для задач, где важно не допускать больших ошибок и при этом измерять среднюю ошибку от истинного результата, MAE – для интерпретации в исходных единицах измерения и нейтрализации выбросов в данных, R^2 – для понимания доли объясненной моделью дисперсии в данных, а MSE – для подчеркивания больших ошибок без необходимости интерпретации результата.

Использование данной модели предполагает, что критическая температура сверхпроводимости может быть описана лишь численными химическими и физическими свойствами материала. Это допущение несколько снижает точность предсказаний модели, но вместе с тем, основывается на том, что с числовыми данными работать наиболее удобно, и их проще всего обрабатывать. Благодаря этому модель будет защищена от переобучения, то есть от поиска закономерностей, которые существуют лишь в обучающей выборке.

Подход к решению задачи регрессии заключается в поиске закономерностей в массиве данных, который содержит множество объектов и достаточно много признаков у каждого объекта. В процессе решения задачи выделяются наиболее значимые наборы признаков, которые наилучшим образом прогнозируют критическую температуру и обладают наибольшей информативностью. Это является важным побочным результатом исследования, который представляет значительный

интерес для исследователей, которые занимаются составлением теории сверхпроводимости [11]. Человек бывает неспособен заметить закономерности, особенно если он не ожидает их увидеть, или если законы их действия слишком сложны для того, чтобы их заметить. Машинное обучение не страдает от такой проблемы и может эффективно находить закономерности и выделять важные признаки в данных.

Итак, в данной главе мы рассмотрели необходимость в использовании методов машинного обучения в научных исследованиях, определили классификацию типов задач машинного обучения и отнесли задачу, решаемую в данной работе, к одному из типов задач. Кроме того, в главе был с математической точки зрения определен функционал ошибок, позволяющий измерить качество предсказаний модели на данных.

В следующей главе будет произведен анализ собранных и обработанных данных для построения первичных гипотез о влиянии признаков на величину критической температуры. Также будет осуществлен переход от теории к практике, а именно будут рассмотрены несколько моделей машинного обучения, способных решить задачу регрессии.

ГЛАВА 3. АНАЛИЗ ДАННЫХ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Данная глава содержит переход от подготовки данных и обоснования использования машинного обучения к выдвижению статистических гипотез на основе данных, а также проверки этих гипотез с помощью применения различных моделей машинного обучения. В параграфе 3.1 производится анализ данных, позволяющий выдвигать первичные гипотезы о важности признаков и их влиянии на значение критической температуры сверхпроводимости. Параграф 3.2 описывает работу с базовой моделью регрессии, с результатами которой в дальнейшем сравниваются результаты других более сложных моделей. Параграф 3.3 описывает работу с моделью «случайный лес» (Random Forest). И параграф 3.4 содержит информацию о модели градиентного бустинга XGBoost.

3.1. Анализ собранных данных

Перед построением различных моделей машинного обучения хорошей идеей является анализ данных. Это позволяет обратить внимание исследователя на некоторые особенности данных (например, дисбаланс целевых значений в обучающей выборке в определенную сторону или сильная корреляция каких-либо признаков с целевыми значениями), а также выдвинуть первичные гипотезы о важности и влиянии признаков на целевые значения. Затем эти гипотезы эффективно проверяются моделями машинного обучения.

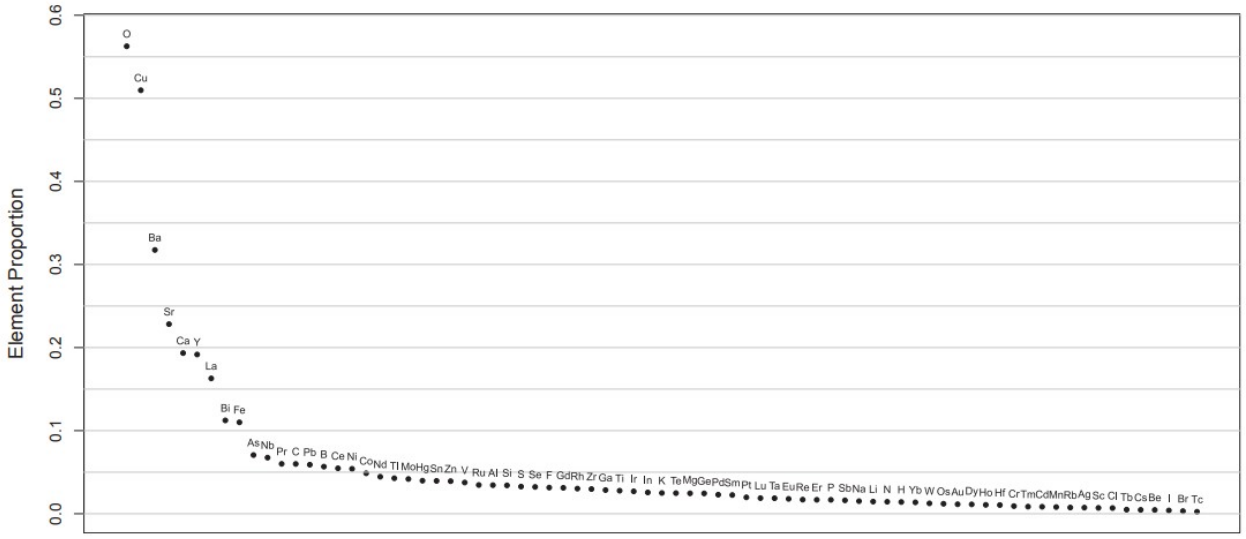

Рис.3.1. Распространенность химических элементов в сверхпроводниках

На рис.3.1 показана доля сверхпроводников, содержащих определенный элемент. Например, кислород присутствует примерно в 56% сверхпроводников. Следующими по распространенности элементами являются медь, барий, стронций и кальций.

Сверхпроводники, содержащие железо, а также содержащие медь (купраты) представляют особый интерес для многих исследовательских задач, поэтому в табл.3.1 приведены некоторые сводные статистические данные о таких элементах. Столбец «Size» показывает число таких материалов из всех собранных данных о 21 263 материалах. Например, 2 339 из 21 263 материалов содержат железо. Остальные столбцы таблицы являются сводной статистикой по наблюдаемым критическим температурам (в Кельвинах): min = минимум, Q1 = первый квартиль, Median = медианное значение, Q3 = третий квартиль, Max = максимум, Mean - среднее значение и SD = стандартное отклонение.

Таблица 3.1

Характеристики сверхпроводников, содержащих железо или медь

Materia	**Size**	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
Iron	2339	0.02	11.3	21.7	35.5	130.0	26.9	21.4
Non-Iron	18924	0.0002	4.8	19.6	68.0	185.0	35.4	35.4
Cuprate	10532	0.001	31.0	63.1	86.0	143	59.9	31.2
Non-Cuprate	10731	0.0002	2.5	5.7	12.2	185	9.5	10.7

На основе этих данных можно построить следующую гипотезу: содержание железа и меди влияет в материале влияет на изменение критической температуры сверхпроводника в сторону увеличения. Причем у купратов критические температуры выше, чем у материалов с содержанием железа.

Железо присутствует примерно в 11% сверхпроводников. Среднее значение T_c сверхпроводников с железом составляет 26,9 ± 21,4 К. Среднее значение для сверхпроводников, не содержащих железа, составляет 35,4 ± 35,4 К; среднее значение и стандартное отклонение оказались одинаковыми. Таким образом, 95%-ный доверительный интервал, основанный на t-распределении, позволяет предположить, что средняя температура сверхпроводников, содержащих железо, ниже, чем у сверхпроводников, не содержащих железа, на 7,4 – 9,5 К. Купраты составляют приблизительно 49,5% сверхпроводников. Среднее значение T_c для купратов составляет 59,9 ± 31,2 К. Таким образом, 95%-ный доверительный интервал, основанный на t-распределении, указывает на то, что среднее значение T_c для купратов выше, чем среднее значение T_c для некупратов, на 49,8 – 51,0 К.

График на рис.3.2 показывает гистограмму распределения всех значений T_c. Значения смещены вправо с достаточно крупным отклонением около $T_c = 80$ К.

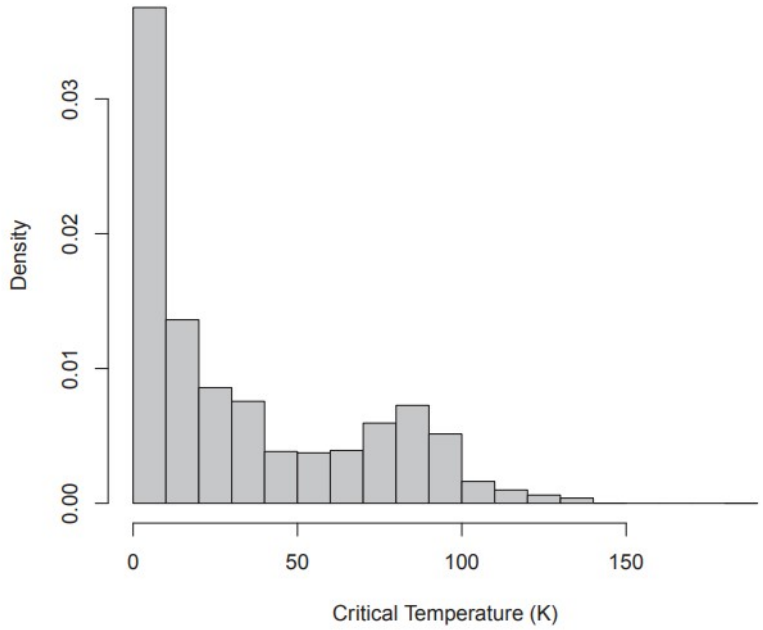

Рис.3.2. Распределение критической температуры сверхпроводников из базы данных

Кроме того, табл.3.2 показывает сводные статистические данные о всех 21 263 элементах, находящихся в базе данных.

Таблица 3.2

Характеристики сверхпроводников, содержащих железо или медь

Min	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
0.00021	5.4	20	63	185.0	34.4	34.2

На рис.3.3 показано среднее значение T_c, сгруппированное по элементам. Ртутьсодержащие сверхпроводники имеют самое высокое значение T_c, в среднем около 80 К. Однако это еще не все. На рис.3.4 показано стандартное отклонение T_c, сгруппированное по элементам. Хотя ртутьсодержащие сверхпроводники в целом имеют самый высокий показатель T_c, эти же материалы демонстрируют четвертую по величине вариабельность (разбросанность) T_c. Фактически, график зависимости среднего значения T_c от стандартного отклонения T_c на рис.3.5 показывает, что в целом, чем выше среднее значение T_c, тем выше вариабельность T_c для каждого элемента.

Эти результаты стоит помнить во время дальнейшей работы с моделями машинного обучения. Чем выше предсказанная критическая температура, тем более она может отличаться от истинной.

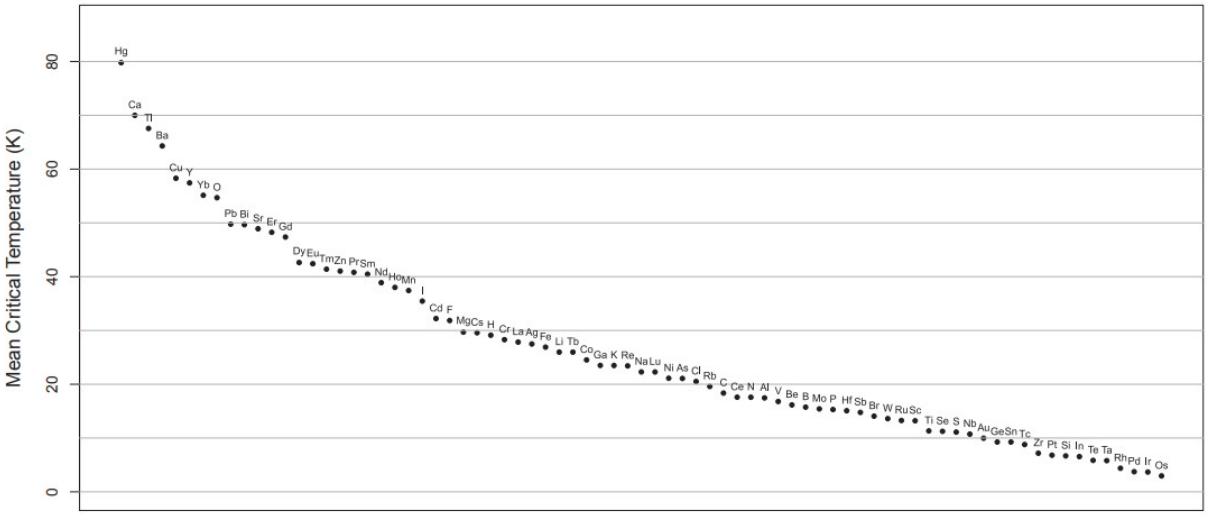

Рис.3.3. Распределение среднего значения критической температуры по элементам

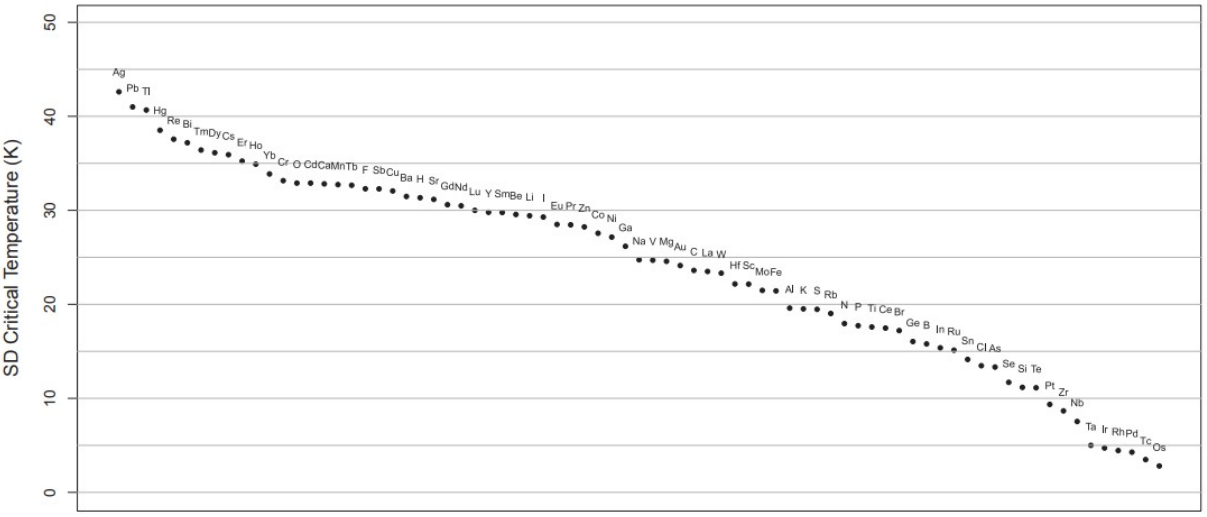

Рис.3.4. Распределение стандартного критической температуры по элементам

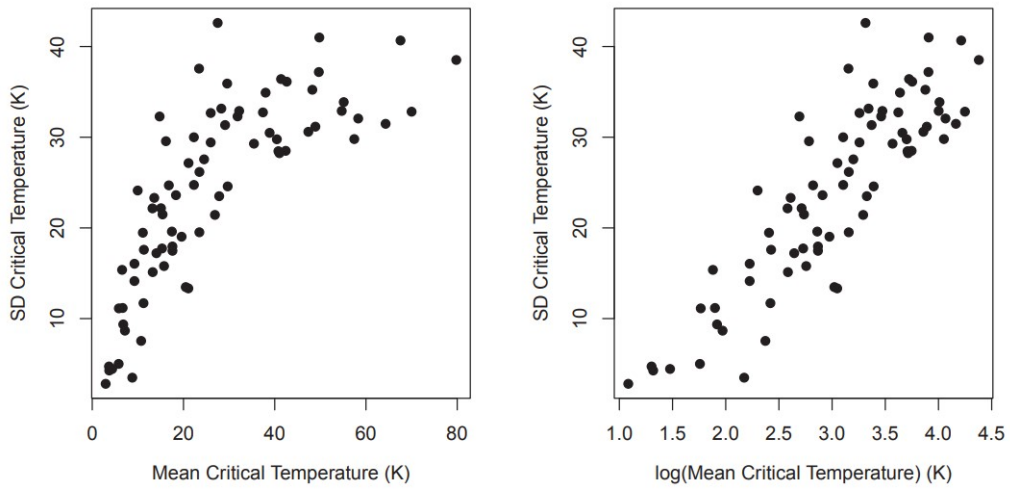

Рис.3.5. Зависимость стандартного отклонения от среднего значения критической температуры по элементам

3.2. Использование базовой модели регрессии

В качестве базовой модели используется сплайн-регрессия с регуляризацией Elastic Net. Сплайн-регрессию можно рассматривать как способ добавления нелинейности к модели линейной регрессии. Она включает в себя подгонку отдельных полиномов низкой степени к различным областям значений объясняющих переменных (признаков). Точки, отмечающие границы различных областей значений, в которые устанавливаются отдельные полиномы, называются узлами. Сплайновая регрессия устанавливает полиномы таким образом, чтобы общая расчетная функция регрессии была непрерывной и гладкой в узлах. Регуляризация Elastic Net добавляется в модель сплайн-регрессии, чтобы сбалансированно уменьшить коэффициенты некоторых признаков с низкой объясняющей способностью до нуля, что снижает вероятность переобучения модели [14].

Регуляризация в машинном обучении — это метод, который используется для предотвращения переобучения модели, улучшения её обобщающей способности и улучшения качества оценки параметров модели. Регуляризация добавляет штрафные значения к функции потерь, что позволяет ограничивать сложность модели и избежать неконтролируемого роста весов для признаков.

Основные виды регуляризации включают L1-регуляризацию (Lasso), L2-регуляризацию (Ridge) и Elastic Net.

L2-регуляризация добавляет штрафное значение, пропорциональное сумме квадратов весов (коэффициентов) модели. Функция потерь с L2-регуляризацией выглядит следующим образом:

$$L(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w} \right)^2 + \lambda \sum_{j=1}^{p} w_j^2 \tag{3.1}$$

, где λ — это параметр регуляризации, контролирующий степень силы влияния штрафного значения, **w** - вектор подобранных моделью коэффициентов w_j для признаков. В случае Ridge-регуляризации, небольшие значения λ позволяют сохранить даже небольшие коэффициенты, не полностью обнуляя их. Таким образом, Ridge-регуляризация помогает выделить объясняющие переменные с малым, но ненулевым вкладом.

L1-регуляризация (Lasso) добавляет штрафное значение, пропорциональное сумме абсолютных значений коэффициентов модели. Функция потерь с

L1-регуляризацией имеет вид:

$$L(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w} \right)^2 + \lambda \sum_{j=1}^{p} |w_j| \tag{3.2}$$

В отличие от L2-регуляризации, L1-регуляризация склонна к обнулению некоторых коэффициентов, тем самым выполняя автоматический отбор признаков. Это означает, что такая регуляризация может выделить более значимые объясняющие переменные, а остальные обнулить.

Не очень строгим, но довольно интуитивным образом это можно объяснить так: в точке оптимума линии уровня регуляризационного члена касаются линий уровня основного лосса (функции потерь), потому что, во-первых, и те, и другие являются выпуклыми, а во-вторых, если они пересекаются трансверсально, то существует более оптимальная точка. Линии уровня L1-нормы – это n-мерные октаэдры. Точки их касания с линиями уровня лосса, скорее всего, лежат на грани размерности, меньшей n-1, то есть как раз в области, где часть коэффициентов-координат равна нулю: как показано на рис.3.6:

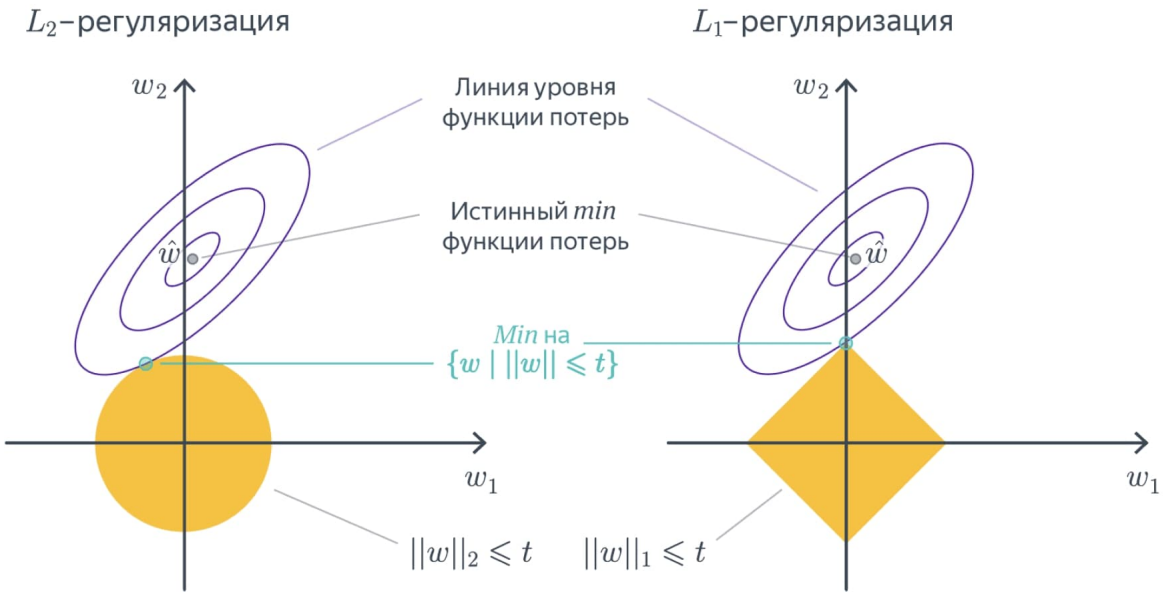

Рис.3.6. Графическое объяснение различия между регуляризациями [14]

Регуляризация **Elastic Net** объединяет L1 и L2 регуляризации и выглядит так:

$$L(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w} \right)^2 + \lambda_1 \sum_{j=1}^{p} |w_j| + \lambda_2 \sum_{j=1}^{p} w_j^2 \tag{3.3}$$

Elastic Net полезна в ситуациях, когда есть сильная корреляция между признаками. Она унаследует свойства как L1, так и L2 регуляризации: может

обнулять некоторые коэффициенты (как Lasso), но сохраняет часть из них малыми, ненулевыми (как Ridge). Это позволяет сбалансированно выделять значимые объясняющие переменные. Поэтому в данной работе используется именно этот тип регуляризации.

Гиперпараметры модели настраиваются до начала процесса обучения модели и определяют ее характеристики. В отличие от обучаемых параметров модели (таких как коэффициенты в линейной регрессии), они не оптимизируются в ходе обучения. Они задаются извне и оказывают влияние на процесс обучения и работы модели. Для достижения наилучшей производительности модели на новых данных часто используются методы перекрестной проверки и поиска гиперпараметров через сетку или случайный подбор значений [14].

Все гиперпараметры модели были выбраны с помощью поиска по сетке. Для осуществления поиска по сетке требуется указать значения гиперпараметров для проверки. Затем модели оцениваются с использованием каждой комбинации значений гиперпараметров и проверяются с помощью перекрестной проверки (Cross Validation). Оптимальные гиперпараметры - это те, которые в нашем случае дают наилучшую по достигнутому результату модель.

Перекрестная проверка (Cross Validation) - это метод, который позволяет оценить эффективность модели на основе данных, которые модель не использовала в процессе обучения, при этом избежав влияния особенностей разбиения данных на результат оценки. Это достигается следующим образом: полный набор данных разбивается на n частей. На каждой итерации модель обучается с использованием данных из $n - 1$ частей, а затем оценивается на оставшейся части данных. Эта процедура повторяется n раз, поэтому каждая часть данных используется один раз в качестве набора для проверки. После этого вычисляется средний показатель оценки. Для всех моделей, представленных далее, используется трехкратная стратифицированная перекрестная проверка. Трехкратная проверка означает, что $n = 3$ в данном случае. Стратифицированная перекрестная проверка означает, что доля категорий целевой переменной одинакова в каждом случае.

Гиперпараметры модели подбираются с использованием класса `GridSearchCV` из библиотеки `sklearn.model_selection` языка программирования `Python`. Этот класс позволяет производить перебор гиперпараметров по заданной сетке с использованием перекрестной проверки.

3.3. Использование модели «случайный лес»

В качестве следующей модели рассматривается модель «случайный лес» (Random Forest), который представляет собой ансамблевый метод, основанный на таком алгоритме машинного обучения как дерево.

Ансамблевые методы в машинном обучении — это подходы, которые объединяют результаты предсказаний нескольких моделей для улучшения общего результата по сравнению с одиночными моделями. Основная идея и преимущество ансамблевых методов заключается в том, что комбинация нескольких моделей позволяет снизить ошибки, а значит увеличить точность предсказаний благодаря учету разнообразия различных моделей.

Основные виды ансамблевых методов:

– **Бэггинг (Bootstrap Aggregating)** предполагает создание множества независимых моделей, обученных на различных выборках данных, полученных с помощью бутстрепинга (случайной выборки с возвращением). Предсказания этих моделей затем усредняются (для регрессии) или берется большинство голосов (для классификации).

  Пример: «Случайный лес» (Random Forest) — ансамбль решающих деревьев, где каждое дерево обучается на случайном подмножестве данных и признаков.

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} h_m(\mathbf{x}) \tag{3.4}$$

  , где $\hat{y}$ - итоговое предсказание, M — количество моделей, h_m — предсказание m-й модели, $\mathbf{x}$ — входные данные.

– **Бустинг (Boosting)** последовательно обучает модели таким образом, что каждая последующая модель исправляет ошибки предыдущих. В результате создается сильный ансамбль моделей, где каждая новая модель фокусируется на сложных для предыдущей модели примерах. Пример: градиентный бустинг (подробнее о градиентном бустинге в параграфе 3.4)

$$\hat{y} = \sum_{m=1}^{M} \alpha_m h_m(\mathbf{x}) \tag{3.5}$$

  , где $\hat{y}$ - итоговое предсказание, α_m — вес m-й модели.

– **Стекинг (Stacking)** предполагает комбинирование предсказаний нескольких базовых моделей с помощью модели-мета (метамодели). Базовые

модели обучаются на исходных данных, а метамодель — на предсказаниях этих базовых моделей. Пример: использование логистической регрессии или другой модели для комбинирования предсказаний нескольких моделей (например, решающих деревьев, линейных моделей и нейронных сетей).

$$\hat{y} = g(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_M(\mathbf{x})) \tag{3.6}$$

, где $\hat{y}$ - итоговое предсказание, g — метамодель, h_m — предсказания базовых моделей.

Преимущества ансамблевых методов:

- **Улучшение точности:** Ансамбли часто обеспечивают более высокую точность по сравнению с отдельными моделями.
- **Снижение риска переобучения:** Комбинирование моделей помогает уменьшить риск переобучения, так как ошибки отдельных моделей могут компенсировать друг друга.
- **Робастность:** Ансамбли более устойчивы к выбросам и шуму в данных.

Дерево решений - это метод классификации/регрессии, который работает путем рекурсивного разделения пространства объектов на две части до тех пор, пока не будет выполнен какой-либо критерий остановки, такой как например максимальная глубина дерева или минимальное число объектов в разделении. Это позволяет получить древовидное представление пространства объектов. Прогнозы делаются путем следования по дереву решений от вершины дерева к нижнему подмножеству, где прогноз делается как среднее значение наблюдений в нижнем подмножестве в случае регрессии и как наиболее часто встречающийся класс в случае классификации.

Модель случайного леса работает, генерируя n бутстрэп выборок из исходного набора данных со случайным набором признаков, взятых в каждой бутстрэп выборке. Бутстрэп выборки формируются путем случайного выбора m наблюдений из исходного набора данных с заменой (то есть с возможностью повторения некоторых наблюдений в сгенерированной выборке), где m - количество наблюдений в исходном обучающем наборе. Затем каждое из деревьев обучается на одной из бутстрэп выборок, и делает собственный прогноз. Наконец, наиболее частый результат, предсказанный отдельными деревьями принятия решений, принимается в качестве окончательного прогноза, сделанного моделью случайного леса.

На рис.3.7 показана визуализация пример работы одного дерева. Дерево решает задачу предсказания зарплаты сотрудника в зависимости от номера его

позиции. Это упрощенный пример всего с одним признаком, но он может дать понимание, как работает дерево решений. «Случайный лес» содержит множество таких деревьев, каждое из которых обучалось на определенной подвыборке исходных данных и признаков.

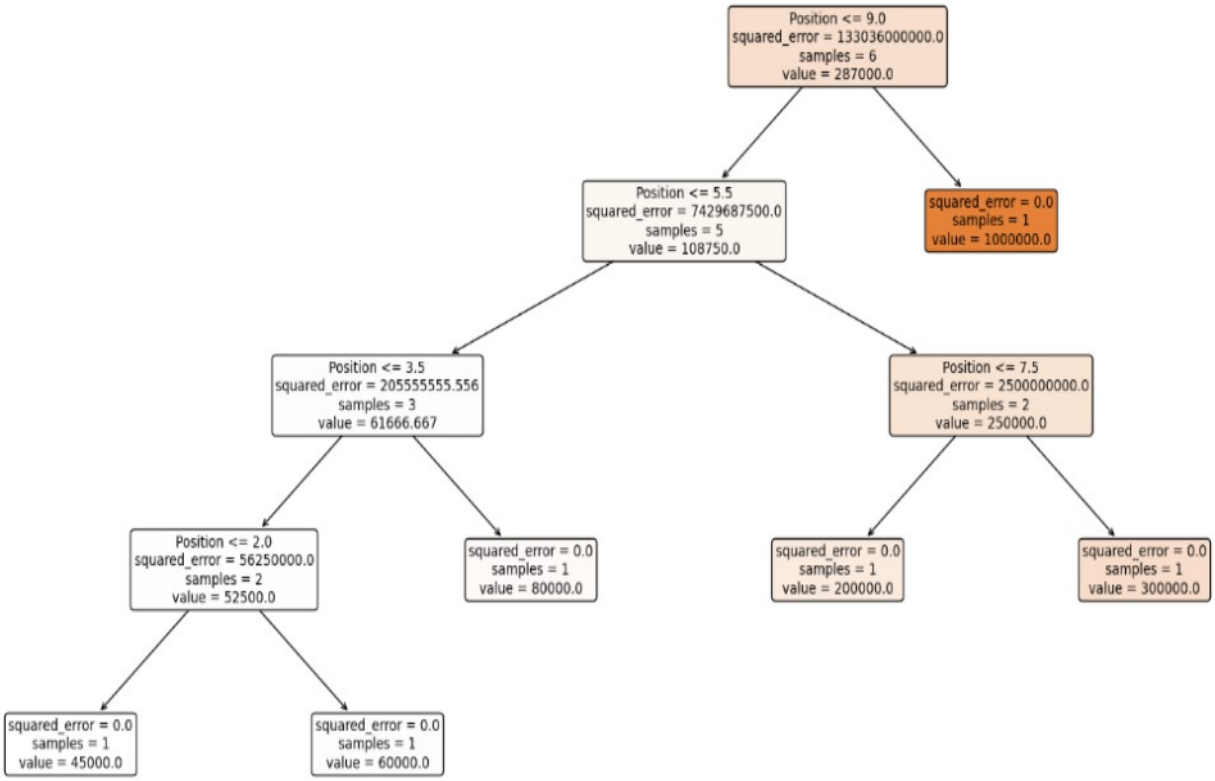

Рис.3.7. Схема работы модели дерева в машинном обучении [17]

3.4. Использование модели градиентного бустинга

Градиентный бустинг - это еще один ансамблевый метод, основанный на алгоритме дерева. В отличие от модели случайного леса, которая обучает отдельные деревья на нескольких бутстрэп выборках параллельно, градиентный бустинг обучает деревья последовательно, причем каждое дерево стремится найти закономерности в той части вариации данных, которая не объясняется предыдущими деревьями. Прогнозы, полученные с помощью модели дерева регрессии с градиентным бустингом, можно рассматривать как взвешенную линейную комбинацию прогнозов дерева регрессии, обученного на исходном наборе данных, плюс подгонки дерева регрессии с использованием ошибок из первого дерева в качестве признаков, плюс подгонки дерева регрессии с использованием ошибок из второго дерева в качестве признаков, и так далее.

Градиентный бустинг называется так, потому что при обучении моделей бустинга используется метод градиентного спуска для минимизации функции потерь. Рассмотрим это подробнее.

Пусть имеется обучающая выборка $\{(x_i, y_i)\}_{i=1}^{N}$, где x_i — признаки, а y_i — целевые значения. Требуется найти модель $F(x)$, которая приближает целевую функцию.

Модель градиентного бустинга строится итеративно. На каждой итерации происходит корректировка текущей модели $F_{m-1}(x)$, чтобы уменьшить ошибку. Таким образом, модель обновляется по следующему правилу:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x), \tag{3.7}$$

, где ν — коэффициент шага (как правило, его называют learning rate), который контролирует интенсивность обновления модели.

Для того чтобы определить $h_m(x)$, градиентный бустинг минимизирует функцию потерь $L(y, F(x))$. Например, при использовании MSE (Mean Squared Error) функция потерь имеет вид:

$$L(y, F(x)) = \frac{1}{2}\sum_{i=1}^{N}(y_i - F(x_i))^2. \tag{3.8}$$

Для нахождения $h_m(x)$ используется градиентный спуск. Градиент функции потерь по $F(x)$ определяет направление максимального увеличения функции потерь, то есть нарастания ошибки. Поэтому отрицательный градиент указывает на направление её наискорейшего уменьшения. Градиент функции потерь для i-го примера на m-ой итерации:

$$g_{im} = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}. \tag{3.9}$$

Для MSE это будет:

$$g_{im} = \left[\frac{\partial \frac{1}{2}(y_i - F(x_i))^2}{\partial F(x_i)}\right]_{F=F_{m-1}} = -(y_i - F_{m-1}(x_i)). \tag{3.10}$$

Следовательно, новый базовый алгоритм $h_m(x)$ должен приближать этот градиент:

$$h_m(x) \approx -g_{im} = y_i - F_{m-1}(x_i). \tag{3.11}$$

Таким образом, на каждой итерации m градиентный бустинг подбирает $h_m(x)$ так, чтобы оно минимизировало ошибку:

$$h_m = \arg\min_h \sum_{i=1}^{N} \left(-g_{im} - h(x_i)\right)^2 . \tag{3.12}$$

Итоговая модель после M итераций выглядит так:

$$F_M(x) = F_0(x) + \nu \sum_{m=1}^{M} h_m(x), \tag{3.13}$$

где $F_0(x)$ — начальная модель (в качестве начальной модели может быть использована константа).

Для реализации модели дерева регрессии с градиентным бустингом существует несколько алгоритмов, таких как XGBoost, CatBoost, LightGBM. В данной работе используется алгоритм XGBoost. Это эффективный алгоритм с поддержкой регуляризации для уменьшения переобучения модели. Оптимальные гиперпараметры для этой модели были выбраны с помощью рандомизированного поиска, который аналогичен ранее описанному поиску по сетке, с той лишь разницей, что рассматривается случайное подмножество комбинаций значений гиперпараметров, а не весь набор всех комбинаций предоставленных значений гиперпараметров. Хотя подход рандомизированного поиска не гарантирует, что будет найдена оптимальная комбинация значений гиперпараметров из всех возможных комбинаций, его использование здесь было обусловлено тем фактом, что алгоритм XGBoost имеет множество гиперпараметров, которые необходимо настроить, что приводит к большому количеству комбинаций значений гиперпараметров, что делает тщательный поиск по сетке невозможным из-за нехватки времени. Поэтому рандомизированный поиск выглядит хорошим подходом, соблюдающим баланс между временем работы кода и оптимизацией гиперпараметров по большой сетке.

Отдельно стоит отметить, что масштабируемость и скорость - важные преимущества использования бустинга XGBoost по сравнению с моделью «случайный лес» [3].

На рис.3.8 показана принципиальная визуализация пример работы градиентного бустинга. Обучение композиции можно представить как перемещение предсказания из точки $(a_k(x_1), a_k(x_2), ..., a_k(x_N))$ в точку $(a_{k+1}(x_1), a_{k+1}(x_2), ..., a_{k+1}(x_N))$. В конечном итоге ожидается, что точка $(a_K(x_1), a_K(x_2), ..., a_K(x_N))$ будет располагаться как можно ближе к точке с истин-

ными значениями $(y_1, y_2, ..., y_N)$. В данном случае точки $(a_i(x_1), a_i(x_2), ..., a_i(x_N))$ означают результат прогнозирования бустинга после применения i моделей в ансамбле.

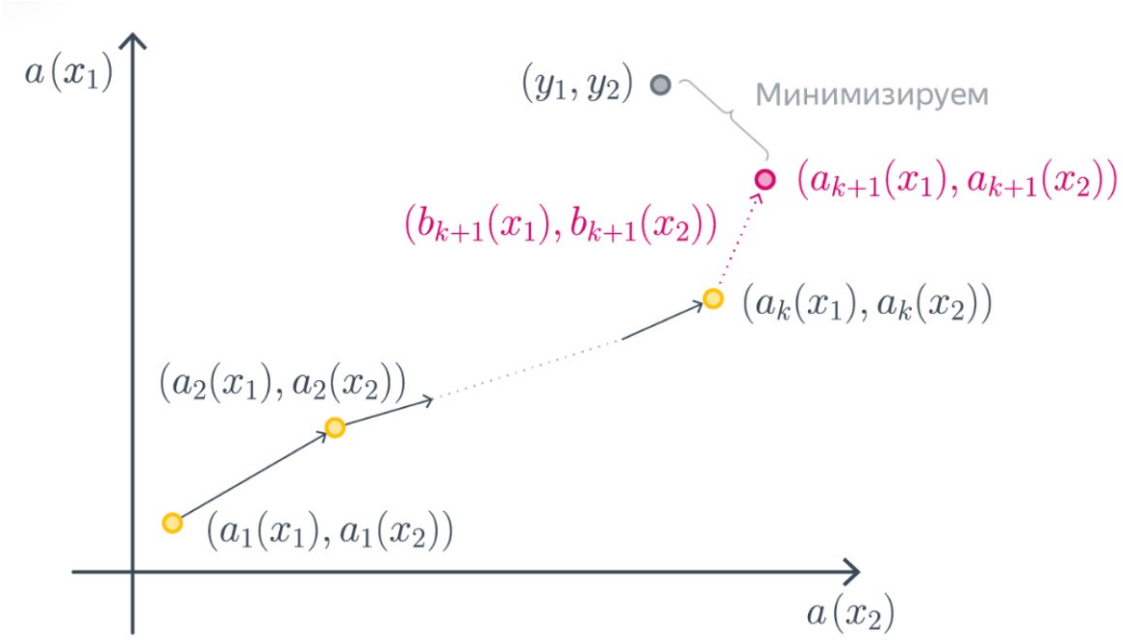

Рис.3.8. Принципиальная схема работы градиентного бустинга в машинном обучении [13]

Таким образом, в данной главе был проведен анализ данных, выдвинуты первичные гипотезы о влиянии признаков на критическую температуру. Кроме того, были построены и специфицированы с точки зрения математики три модели машинного обучения - линейная сплайн-регрессия, «случайный лес» и градиентный бустинг.

В следующей главе будет проведен анализ результатов работы этих моделей и выбрана лучшая модель для решения данной задачи.

ГЛАВА 4. АНАЛИЗ РЕЗУЛЬТАТОВ МОДЕЛЕЙ

В данной главе осуществляется анализ достигнутых моделями машинного обучения результатов. В параграфе 4.1 приведено описание процесса подбора оптимальных гиперпараметров для каждой из моделей. В параграфе 4.2 проводится сравнение эффективности данных моделей машинного обучения по метрике качества, определенной ранее. Параграф 4.3 содержит анализ важности признаков в данных в определении критической температуры сверхпроводимости.

4.1. Оценка качества работы различных моделей

Процесс оценки качества моделей происходит следующим образом:

A. Собранные данные случайным образом разделяются на n равных частей для перекрестной проверки. Одна из частей используется для оценки качества, остальные $n - 1$ частей - для обучения модели;
B. Модель с нуля учится на обучающих данных;
C. На тестовых данных прогнозируется значение T_c;
D. На основе прогнозируемых значений по тестовым данным и реальных значений T_c элементов из тестовых данных рассчитывается метрика качества RMSE;
E. Предыдущие три шага повторяются n раз со сменой тестовой выборки;
F. Затем n полученных результатов RMSE усредняются.

Полученное таким образом значение RMSE и считается итоговой метрикой качества модели. Эту метрику легко интерпретировать - она показывает, на сколько градусов в среднем ошибается модель в прогнозировании T_c.

Далее рассматриваются оптимальные подобранные гиперпараметры и результаты RMSE для трех различных моделей, построенных в предыдущей главе.

4.1.1. Сплайн-регрессия с регуляризацией

Используется поиск по сетке (`GridSearchCV`) для нахождения оптимального набора двух гиперпараметров:

– количество узлов для создания сплайнов `spline_n_knots` среди возможных значений `[15, 20]`. Увеличение числа узлов позволяет создавать больше сплайнов и точнее приближать нелинейность искомой функции[18];

– коэффициент регуляризации `ridge_alpha` среди возможных значений в интервале от -1 до 10, равномерно разбитом на 10 узлов. Чем больше коэффициент регуляризации, тем выше штраф модели за большие веса у признаков;

По результатам подбора гиперпараметров наилучший результат на тестовой выборке показала модель с гиперпараметрами `spline_n_knots=20`, `ridge_alpha=1.44`.

Ее метрика RMSE на перекрестной проверке равна **13.59 К**.

4.1.2. «Случайный лес»

Здесь также используется поиск по сетке (`GridSearchCV`) для нахождения оптимального набора двух гиперпараметров:

– доля признаков для расщепления в каждом из деревьев (признаки выбираются случайно) `max_features` среди возможных значений `[0.2, 0.3, 0.5]`. При увеличении `max_features` увеличивается время построения леса, а деревья становятся «более однообразными»[12];
– максимальная глубина деревьев `max_depth` среди возможных значений `[10, 20, 50]` Увеличение глубины увеличивает время построения леса, но так же способно точнее предсказывать результат, если данные не шумные (не содержат много выбросов)[12].

Количество деревьев в случайном лесу в данной работе зафиксировано на уровне 750, а максимальная доля признаков, используемых для обучения каждого дерева, и максимальная глубина каждого дерева были определены с помощью поиска по сетке с 5-кратной перекрестной проверкой. В выбранной модели случайного леса с наилучшими гиперпараметрами каждое дерево имеет максимальную глубину `max_depth` в 20 узлов и обучается на 50% объектов, выбранных случайным образом (`max_features`).

Модель случайного леса работает лучше, чем сплайн-регрессия, при перекрестной проверке RMSE составляет около **9.06 К**.

4.1.3. Градиентный бустинг XGBoost

В отличие от двух предыдущих подборов гиперпараметров, здесь используется не поиск по сетке, а случайный поиск `RandomizedSearchCV`. Это необходимо для того, чтобы ускорить время подбора гиперпараметров, при этом не теряя сильно в качестве.

- число моделей деревьев в бустинге `n_estimators` с заданным значением `[750]`;
- максимальная глубина деревьев `max_depth` среди возможных значений `[10, 20]`;
- `min_child_weight` среди возможных значений `[5, 10]`. Минимальное количество объектов, необходимое для формирования конечного узла (конца ветви дерева). Ограничивает возможность дерева создавать дочерние узлы с малым числом объектов. Благодаря этому ограничению может повышаться обобщающая способность модели.
- темп обучения `learning_rate` среди возможных значений `[0.01, 0.02]`. Параметр управляет величиной изменений, допустимых при переходе от одного дерева к другому;
- доля выборки `subsample` с заданным значением `[0.75]`. Величина определяет, какая часть исходного набора данных попадает случайным образом в выборку во время каждой итерации процесса обучения;
- доля объектов для обучения дерева (объекты выбираются случайным образом) `colsample_bytree` среди возможных значений `[0.2, 0.3]`;
- коэффициент при L1-регуляризации `alpha` среди возможных значений `[0, 1, 2]`;
- коэффициент при L2-регуляризации `lambda` среди возможных значений `[0, 1, 10]`;
- параметр `gamma` среди возможных значений `[2, 5, 10]`. Параметр используется для контроля склонности модели к переобучению. Так как модели деревьев склонны к переобучению, важно подбирать правильную величину этого параметра.

Выбранная по результатам подбора гиперпараметров [21] модель XGBoost имеет 750 базовых моделей (`n_estimators=750`), каждая из которых обучается на отдельных случайных подвыборках, содержащих 75% наблюдений и 30% признаков. Каждое дерево выращивается на максимальную глубину 20, и выбранные гиперпараметры регуляризации указывают на то, что регуляризация в модели является относительно сильной. Модель XGBoost немного превосходит модель случайного леса. Перекрестная проверка RMSE составляет для градиентного бустинга около **8.89 K**.

4.2. Анализ достигнутых моделями результатов

Таким образом, по результатам подбора гиперпараметров в трех моделях - сплайн-регрессии с регуляризацией, «случайного леса» и градиентного бустинга на примере XGBoost, получены значения метрики RMSE с использованием наилучших подобранных гиперпараметров, показанные в табл.4.1.

Таблица 4.1

Результаты оценки качества моделей машинного обучения по метрике RMSE

Рейтинг	**Модель**	**RMSE**
1	XGBoost	8.89
2	«Случайный лес»	9.06
3	Сплайн-регрессия с регуляризацией	13.59

На рис.4.1 показано сравнение предсказанной с помощью XGBoost критической температуры с наблюдаемой в реальности. На этом графике сильного смещения прогнозов от реальности не наблюдается. Есть только несколько заметных выбросов, которые, по видимому, плохо поддаются общему прогнозированию, и возможно зависят от каких-то еще признаков, не учтенных в текущем наборе данных.

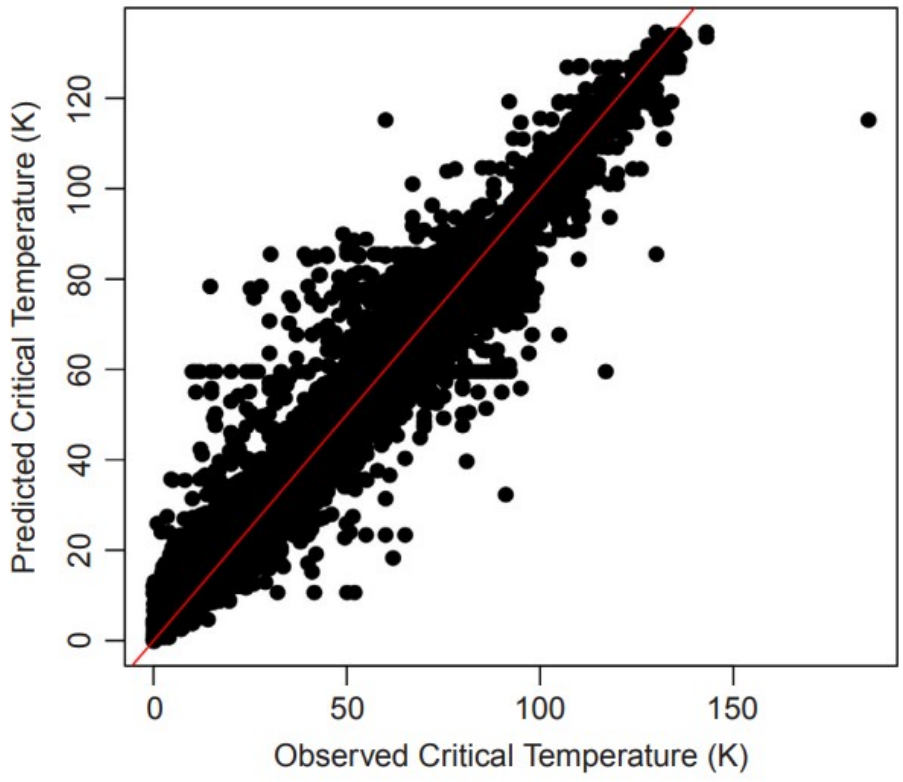

Рис.4.1. Сравнение прогнозируемой XGBoost и реальной критической температуры

4.3. Определение важности признаков

Стоит также уделить внимание вопросу выявления важности признаков и их вкладу в определение критической температуры. Важность характеристик в XGBoost измеряется коэффициентом Gain. Коэффициент Gain для признака определяется следующим образом: Каждый раз, когда узел дерева разбивается по данному признаку, фиксируется улучшение целевой функции. Коэффициент Gain для признака вычисляется по формуле refeq:gain.

$$\text{Gain для признака} = \frac{\text{Сумма всех Gain для признака}}{\text{Сумма всех Gain для } \textbf{всех} \text{ признаков}} \tag{4.1}$$

Признаки с более высоким коэффициентом Gain являются более важными.

Таблица 4.2

Список 20 наиболее важных признаков для прогнозирования T_c, составленных на основе свойств элементов, составляющих сверхпроводник

Признак	**Значение Gain**
Разность теплопроводностей	0.295
Взвешенное станд. отклон. теплопроводностей	0.084
Разность атомных радиусов	0.072
Взвешенное геом. среднее теплопроводностей	0.047
Станд. отклон. теплопроводностей	0.042
Взвешенная энтропия валентностей	0.038
Взвешенное станд. отклон. сходства к электрону	0.036
Взвешенная энтропия атомных масс	0.025
Взвешенное среднее валентностей	0.022
Взвешенное геом. среднее сходства к электрону	0.021
Взвешенная разность сходства к электрону	0.016
Взвешенное среднее теплопроводностей	0.015
Взвешенное геом. среднее валентностей	0.014
Станд. отклон атомных масс	0.013
Станд. отклон. плотностей	0.010
Взвешенная энтропия теплопроводностей	0.010
Взвешенная разность теплопроводностей	0.010
Взвешенное среднее атомных масс	0.009
Взвешенное станд. отклон. атомных масс	0.009
Среднее геометрическое плотностей	0.009

В табл.4.2 представлены 20 наиболее важных признаков. Характеристики, полученные на основе теплопроводности, атомного радиуса, валентности, сродства к электрону и атомной массы представляются наиболее важными. Также стоит

обратить внимание, что признаки, определенные на основе указанных выше химических свойств, появляются в списке чаще всего. Это говорит о том, что эти свойства являются более важными, чем другие при предсказании T_c.

Таким образом, можно сказать, что наиболее важным свойством для определения T_c является коэффициент теплопроводности, различные статистики которого входят в топ-5 признаков целых 4 раза.

ЗАКЛЮЧЕНИЕ

В ходе данной работы была реализована модель машинного обучения для предсказания критической температуры сверхпроводника на основе его химических и физических свойств. Она позволяет на основе данных о формуле сверхпроводника подготовить признаки и прогнозировать значение критической температуры данного сверхпроводника. Кроме того, данная модель выделяет информацию о важности определенных свойств материала, а также их статистических мерах для определения критической температуры T_c. Благодаря этому становится возможным выделить свойства сверхпроводников, на которые исследователям стоит обратить особое внимание при создании теории сверхпроводимости и синтезе новых сверхпроводников.

Был проведен сбор данных из нескольких источников, объединение и подготовка этих данных к использованию в модели машинного обучения. Затем на основе обработанных и очищенных от ошибок и выбросов данных были подготовлены признаки, включающие в себя несколько статистик химических свойств для каждого сверхпроводника, таких как среднее арифметическое, среднее геометрическое, энтропия, и др.

Затем был определен подход для реализации модели, выяснены предпосылки к использованию методов машинного обучения в данной задаче. После этого были рассмотрены типы задач в терминах методов машинного обучения, и данная задача была отнесена к одному из рассматриваемых типов, а именно к задаче регрессии. Таким образом стало возможно поставить задачу на математическом языке, а также определить критерии оценки качества работы моделей машинного обучения - RMSE (Root Mean Square Error).

Далее был проведен анализ подготовленных данных с целью определения видимых закономерностей во влиянии признаков на значение критической температуры сверхпроводимости, а также постановки базовых гипотез для дальнейшей проверки их с помощью методов машинного обучения. Также были на принципиальном и математическом уровне рассмотрены три модели машинного обучения, способные решить задачу регресии - базовая модель сплайн-регрессии с регуляризацией, модель «случайный лес», основанная на более простой модели дерева, а также модель градиентного бустинга на примере модели XGBoost.

После этого был проведен подбор гиперпараметров моделей машинного обучения и измерена метрика качества на моделях с лучшими подобранными

гиперпараметрами. Наилучший результат в прогнозировании T_c показала модель градиентного бустинга XGBoost со значением средней ошибки в 8.89 Кельвинов. Следом за ней по качеству оказалась модель «случайный лес» с ненамного меньшей средней ошибкой в 9.06 Кельвинов. Куда более посредственный результат показала модель сплайн-регрессии с регуляризацией - ее средняя ошибка составила 13.59 Кельвинов.

В завершение работы была оценена важность признаков в задаче прогнозирования критической температуры сверхпроводимости. По результатам оценки, признаки, извлеченные на основе теплопроводности, атомного радиуса, атомной массы, валентности и сродства к электрону материала сверхпроводника вносят наибольший вклад в точность предсказания модели.

Данная работа может быть полезна в работе исследователям-практикам, занимающимся проведением экспериментов над сверхпроводниками с целью определения их критической температуры, а также исследователям-теоретикам, занимающимся составлением теории сверхпроводимости и вопросами поиска новых сверхпроводящих материалов.

Таким образом, можно сделать вывод, что цель выпускной квалификационной работы была выполнена и поставленные задачи решены полностью. Построенная модель машинного обучения может применяться в исследовательских и индустриальных целях и послужить инструментом для выбора оптимальных материалов сверхпроводников, уменьшив риски, возникающие вследствие неправильно подобранного метода поиска.

Работа может получить дальнейшее развитие благодаря добавлению в базу данных большего числа признаков сверхпроводников, более умелой работе с признаками с целью извлечения лишь важных для работы модели, а также с внедрением более сложных архитектур моделей машинного обучения таких как например нейронные сети.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Шмидт В. Введение в физику сверхпроводников. — 2-е изд., испр. и доп. — Москва: Изд-во МЦНМО, 2000. — С. 3—14. — (Сер.: Современные лекционные курсы).
2. Белоусов О., Палий Н. К вопросу о критической температуре сверхпроводимости // Russian Metallurgy (Metally). — 2012. — С. 16—34.
3. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. — 2016.
4. Conder K. A second life of the Matthias's rules // Superconductor Science and Technology. — 2016. — Т. 29. — С. 080502.
5. Dick J. Calculation of the relative metastabilities of proteins using the CHNOSZ software package // Geochemical transactions. — 2008. — Т. 9. — С. 10.
6. Hassenzahl W. Applications of superconductivity to electric power systems // Power Engineering Review, IEEE. — 2000. — Т. 20. — С. 4—7.
7. Machine learning modeling of superconducting critical temperature / V. Stanev [et al.] // Computational Materials. — 2017. — Vol. 4.
8. Matthias B. T. Empirical Relation between Superconductivity and the Number of Valence Electrons per Atom // Phys. Rev. — 1955. — Т. 97, вып. 1. — С. 74—76.
9. Owolabi T., Akande K., Olatunji S. Estimation of Superconducting Transition Temperature TC for Superconductors of the Doped MgB2 System from the Crystal Lattice Parameters Using Support Vector Regression // Journal of Superconductivity and Novel Magnetism. — 2014. — Т. 28.
10. Owolabi T., Akande K., Olatunji S. Prediction of Superconducting Transition Temperatures for Fe-Based Superconductors using Support Vector Machine // Advances in Physics Theories and Applications. — 2014. — Т. 35. — С. 12—26.
11. Больше не нужно вслепую бродить по таблице Менделеева / Наука и жизнь. — 2021. — URL: https://www.nkj.ru/open/42373/ (дата обращения: 08.05.2024).
12. Случайный лес (Random Forest) / КвазиНаучный блог Александра Дьяконова. — 2016. — URL: https://alexanderdyakonov.wordpress.com/2016/11/14/%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/ (дата обращения: 14.05.2024).

13. Учебник по машинному обучению. Градиентный бустинг / Яндекс. — 2023. — URL: https://education.yandex.ru/handbook/ml/article/gradientnyj-busting (дата обращения: 14.05.2024).

14. Учебник по машинному обучению. Линейные модели / Яндекс. — 2023. — URL: https://education.yandex.ru/handbook/ml/article/linear-models (дата обращения: 10.05.2024).

15. ElementData / Wolfram Research. — 2014. — URL: https://reference.wolfram.com/language/ref/ElementData.html (дата обращения: 05.05.2024).

16. Linear Regression Masterclass - ML / Kaggle. — 2023. — URL: https://www.kaggle.com/code/auxeno/linear-regression-masterclass-ml (дата обращения: 08.05.2024).

17. Random Forest Regression in Python / Geeks for geeks. — 2023. — URL: https://www.geeksforgeeks.org/random-forest-regression-in-python/ (дата обращения: 12.05.2024).

18. SplineTransformer / Scikit-learn library documentation. — 2024. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.SplineTransformer.html (дата обращения: 15.05.2024).

19. Superconducting Material Database / Japan's National Institute for Materials Science. — 2016. — URL: http://supercon.nims.go.jp/index_en.html (дата обращения: 06.05.2024).

20. Supervised vs Unsupervised Learning / StrataScratch. — 2020. — URL: https://www.stratascratch.com/blog/supervised-vs-unsupervised-learning/ (дата обращения: 08.05.2024).

21. XGBoost Hyperparameter Tuning - A Visual Guide / Kevin Vecmanis. — 2019. — URL: https://kevinvecmanis.io/machine%20learning/hyperparameter%20tuning/dataviz/python/2019/05/11/XGBoost-Tuning-Visual-Guide.html (дата обращения: 15.05.2024).