МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ

ИНФОРМАТИКА И КИБЕРНЕТИКА

4 (18)

Донецк – 2019

УДК 004.3+004.9+004.2+51.7+519.6+519.7

ИНФОРМАТИКА И КИБЕРНЕТИКА, № 4 (18), 2019,
Донецк, ДонНТУ.

Представлены материалы по вопросам приоритетных направлений научно-технического обеспечения в области информатики, кибернетики и вычислительной техники.

Материалы предназначены для специалистов народного хозяйства, ученых, преподавателей, аспирантов и студентов высших учебных заведений.

Редакционная коллегия
Главный редактор: *Павлыш В. Н., д.т.н., проф.*
Зам. глав. ред.: *Мальчева Р. В., к.т.н., доц.*
Ответственный секретарь: *Воронова А. И.*
Члены редакционной коллегии: *Аверин Г. В., д.т.н., проф.; Аноприенко А. Я., к.т.н., проф.; Зинченко Ю. Е., к.т.н., доц.; Зори С. А., д.т.н., доц.; Карабчевский В. В., к.т.н., доц.; Миненко А. С., д.ф-м.н., проф.; Привалов М. В., к.т.н., доц.; Скобцов Ю. А., д.т.н., проф.; Федяев О. И., к.т.н., доц.; Шелепов В. Ю., д.ф-м.н., проф.*

Рекомендовано к печати ученым советом ГОУ ВПО «Донецкий национальный технический университет» Министерства образования и науки ДНР. Протокол № 9 от 27 декабря 2019.

Свидетельство о регистрации СМИ: серия ААА № 000145 от 20.06.2017.

Приказ МОН ДНР № 135 от 01.02.2019 о включении в Перечень рецензируемых научных изданий ВАК ДНР.

Контактный адрес редакции
ДНР, 83001, г. Донецк, ул. Артема, 58, ГОУ ВПО «ДонНТУ»,
4-й учебный корпус, к. 36., ул. Кобозева, 17.
Тел.: +38 (062) 301-07-35, +38 (071) 334-89-11
Эл. почта: infcyb.donntu@yandex.ru
Интернет: http://infcyb.donntu.org

© Донецкий национальный технический университет
Министерство образования и науки ДНР, 2019

СОДЕРЖАНИЕ

Информатика и вычислительная техника

УДК 519.237;004.67

Регрессионная модель для прогнозирования температуры вспышки дизельного топлива в закрытом тигле

А. Ю. Максимова*, А. А. Иванова*, Н. С. Лозинский **
*ГУ «Институт прикладной математики и механики», г. Донецк
**ГУ «Институт физико-органической химии и углехимии им. Л.М. Литвиненко», г. Донецк
maximova.alexandra@mail.ru, ivanova.iamm@mail.ru, lozinsky58@mail.ru

Аннотация
Рассматривается задача построения функциональной зависимости температуры вспышки дизельного топлива в закрытом тигле от его измеренных нормативных параметров с использованием подходов машинного обучения. Выполнен анализ сырых данных лаборатории контроля качества методами корреляционного анализа. Разработан метод поиска аномальных объектов, подозрительных на выбросы. Применена гребневая регрессионная модель, построены вероятностные оценки параметров линейной модели. Предложенный подход может быть применен при анализе похожих данных в области нефтехимии.

Введение

Регрессионный анализ позволяет определять скрытые зависимости между информативными показателями качества контролируемого продукта. С другой стороны существуют отрасли, в которых массивы данных контроля качества имеют большой объем и ежесуточно накапливаются, например, нефтехимия. Поэтому, если применить регрессионный анализ к обработке этих массивов, можно получить зависимости между информативными показателями, которые позволят уменьшить трудозатраты на проведение контроля качества.

Для дизельного топлива (ДТ) информативными показателями являются: температура вспышки в закрытом тигле (ТВЗТ) и фракционный состав (ФС) [1]. Поскольку эти показатели контролируют свойства самой легкой фракции ДТ, между ними должна существовать взаимосвязь [2]. Более того, в [3, 4] построены регрессионные модели зависимости между ТВЗТ и ФС для некоторых видов дизельного топлива, производимых в Индии и Бразилии. В [5] приведено уравнение регрессии для зависимости ТВЗТ от плотности дизельного топлива с содержанием серы до 0,5 % массы.

Упомянутые модели построены для стандартных ДТ промышленного производства. Однако на практике случается контролировать образцы ДТ подвергнутые смешению, как топливами различных производителей, так и горючими материалами различных марок.

Сведения о работах, посвященных обработке данных контроля реальных образцов дизельного топлива, подверженных технологическим и природным воздействиям, не найдены, поэтому авторами проведен анализ таких данных, полученных в лаборатории контроля качества нефтепродуктов г. Донецка.

Постановка задачи

Необходимо построить регрессионную модель для вычисления ТВЗТ дизельного топлива по измеренным показателям (ФС и плотность ρ) и оценить пределы ее применимости и достоверность полученного результата.

Вопросами определения факторов, которые влияют на значение температуры вспышки, занимаются специалисты в предметной области. В литературе приводятся формулы расчета ТВЗТ для известных органических соединений по температуре их кипения и эмпирическим коэффициентам [6, 7]. Для смесей известных органических соединений также можно посчитать температуру вспышки при условии, что точно известно соотношение компонентов смеси. При этом средняя квадратичная погрешность расчета по приведенным формулам достигает 10-13 °С. Дизельные топлива являются сложными смесевыми продуктами, состав которых зависит от особенностей технологического процесса нефтеперерабатывающего предприятия, на котором они произведены, и от используемых для их приготовления компонентов. Вопрос взаимосвязи между фракционным составом дизельного топлива и ТВЗТ рассматривался в работах [3, 4].

В [5] приводятся функциональные зависимости ТВЗТ от плотности и/или вязкости. Целью данной работы является исследование показателей, по которым может быть рассчитано значение ТВЗТ, и восстановление параметров функции регрессии для вычисления значения ТВЗТ по измеренным значениям этих показателей.

Процедура сбора качественных обучающих данных при решении задач анализа

данных занимает значительную часть времени и требует соответствующих ресурсов. В данном исследовании используется обучающая выборка, собранная в течение 10 лет, состоящая из 1157 необработанных наблюдений, содержащих ошибки и пропуски, что связано с тем, что данные были собраны без предварительного плана сбора исходной статистической информации, и в условиях накопления грубых ошибок, обусловленных, в том числе, влиянием человеческого фактора. Данная выборка несбалансированна в том смысле, что содержит значительно больше примеров стандартных образцов с ТВЗТ большей 40 °С. Возможности провести содержательный анализ условий, при которых регистрировались наблюдения, нет. Принимая во внимание особенности выборки, возникла задача поиска аномальных образцов, подозрительных на выбросы.

Разведочный анализ данных

Выборка с сырыми данными содержала как числовые измеренные показатели: ТВЗТ, плотность при 20 °С, ФС с шагом 10 %, кроме последнего значения, измеренного для 96 %, так и нечисловые данные: идентификатор пробы, источник забора образца, название производителя, время забора пробы. При этом у многих образцов есть пропущенные данные, например, известна плотность, но отсутствует ФС, или наоборот.

Введем обозначения для числовых признаков и целевой переменной регрессионной модели: $\mathrm{t_{vzt}}$ – ТВЗТ, $\mathrm{t_0}$ – температура кипения, $t_0, t_{10} \ldots, \mathrm{t_{80}}, t_{96}$ – температуры выкипания фракций, ρ – плотность. При этом каждый из этих признаков определяется с нормируемой сходимостью и воспроизводимостью, которая для ТВЗТ составляет 4 °С, для точек ФС – 2-7 °С.

На первом этапе очистим выборку от образов с пропусками и заведомо недопустимыми значениям (например, $t_0 > 500$). После предобработки выборка X^n включает 324 примера ($n = 324$). Оценим тесноту линейной статистической связи между признаками и целевой переменной t_{vzt}, рассчитав значение коэффициента корреляции по формуле Пирсона:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

где x – информативный показатель,
y – целевая переменная t_{vzt},
$\bar{x}, \bar{y}$ – выборочные средние x, y.

Анализ полученных коэффициентов корреляции (табл. 1), рассчитанных как для всей выборки, так и для подвыборки образцов, у которых $t_{vzt} > 40$ в соотвествии с требованиями ДСТУ 3868-99 [8], показал возможную сильную корреляцию между температурой кипения и ТВЗТ $\mathrm{r}_{\mathrm{t_0 t_{tvz}}}^{\text{все}} = 0.77$, $\mathrm{r}_{\mathrm{t_0 t_{tvz}}}^{t_{vzt}>40} = 0.85$, а также возможную среднюю корреляцию между температурой выкипания 10 % фракции и 20 % фракции $\mathrm{r}_{\mathrm{t_{10} t_{tvz}}}^{\text{все}} = 0.69$, $\mathrm{r}_{\mathrm{t_{10} t_{tvz}}}^{t_{vzt}>40} = 0.73$ $\mathrm{r}_{\mathrm{t_{20} t_{tvz}}}^{\text{все}} = 0.58$, $\mathrm{r}_{\mathrm{t_{20} t_{tvz}}}^{t_{vzt}>40} = 0.62$.

Таблица 1 – Коэффициенты корреляции

	t0	**t10**	**t20**	**t30**	**t40**	**t50**	**t60**	**t70**	**t80**	**t90**	**t96**	**ro**
$t_{vzt} > 40$ °С	**0.85**	**0.73**	**0.62**	0.09	0.05	0.39	0.05	0.25	0.16	-0.01	-0.01	0.23
t_{vzt}	**0.77**	**0.69**	**0.58**	0.5	0.43	0.4	0.31	0.25	0.15	0.04	-0.06	0.29

Визуальный анализ корреляционных полей между ТВЗТ и всеми известными показателями (рис. 1) подтверждает наличие достаточно сильной линейной зависимости ТВЗТ от t_0,t_{10},t_{20}.

В [5] приводится регрессионная модель зависимости $\mathrm{t_{vzt}}$ от плотности дизельного топлива с содержанием серы меньше 0,5 % масс

$$\mathrm{t_{vzt}} = 117{,}012\rho_4^{20} - 53{,}944.$$

Поэтому несколько неожиданным оказывается визуально слабо подтвержденная линейная связь между $\mathrm{t_{vzt}}$ и плотностью ρ. Предположительно это может быть связано с достаточно большим числом нестандартных образцов в исследуемой выборке, которые позволяет определить визуальный анализ корреляционных полей.

В результате проведенного разведочного анализа решено строить регрессионную модель с использованием только предикторов t_0, t_{10}, t_{20}.

Выбор регрессионной модели

Множественная линейная регрессия преполагает линейную зависимость исследуемой случайной величины ξ_{tvzt} от предикторов $(x_1, \ldots, x_m)$ при условии, что ее дисперсия σ:

$$\xi_{tvzt} = f(x) + \varepsilon,$$

где $f(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$;

x_1, x_2, x_3 – предикторные переменые, соответсвующие признакам t_0, t_{10}, t_{20};

w_0, w_1, w_2, w_3 – параметры модели;

ε – стохастическая ошибка, которая обусловлена наличием неизвестных факторов, $E(\varepsilon|\boldsymbol{x}) = 0$.

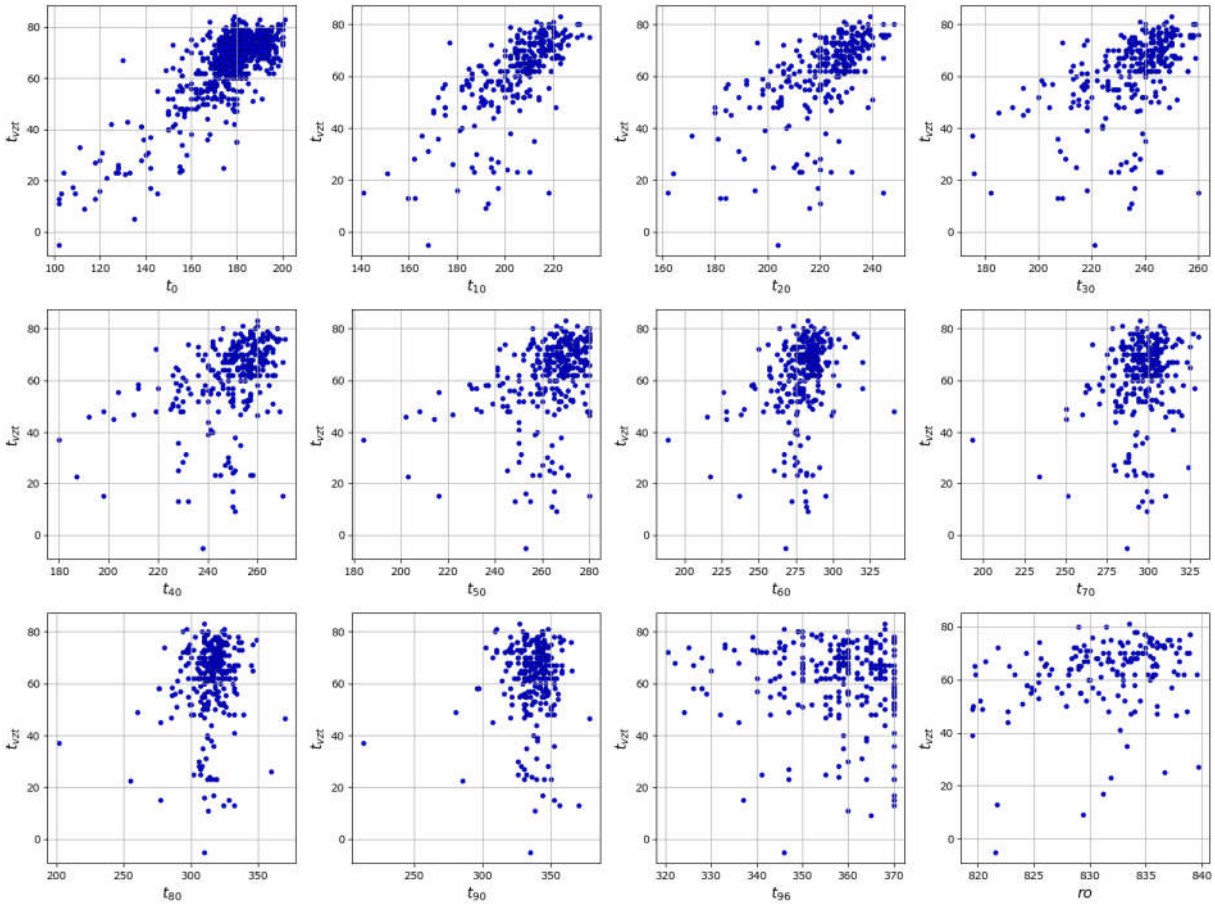

Рисунок 1 – Корреляционные поля полной обучающей выборки

Решим задачу оценки параметров w_0, w_1, w_2, w_3 линейной функции $f(x)$ чтобы вычислить ТВЗТ как

$$y_{tvzt,\text{ср}} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3.$$

Поскольку было установлено, что существенно ТВЗТ зависит только от признаков t_0, t_{10}, t_{20} и остальные признаки исключили из рассмотрения, появилась возможность использовать выборку с большим количеством примеров: 324 образца, включая примеры, где данные по плотности отсутствовали в исходном наборе (рис. 2).

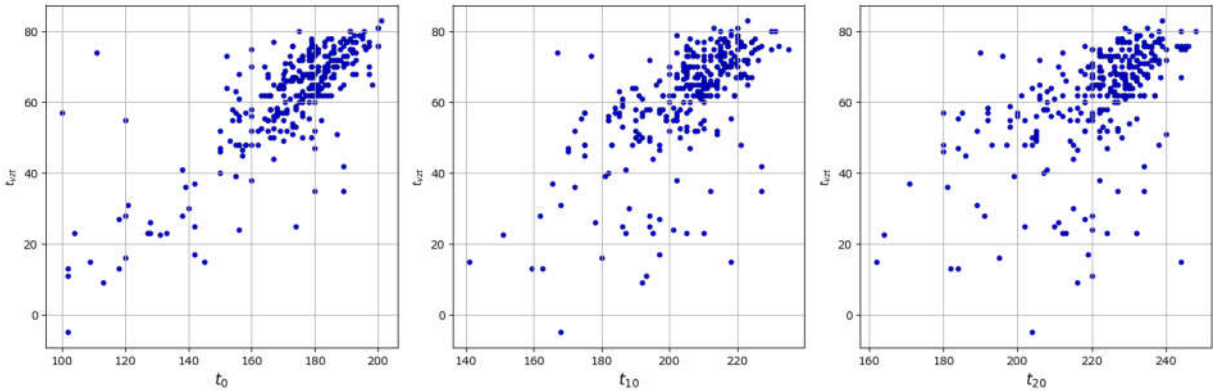

Рисунок 2 – Корреляционные поля

В результате визуального анализа рис. 2 можно предположить, что дисперсия отклонений t_{vzt} от $y_{tvzt,\text{ср}}$ одинакова для всех значения предикторов x.

Настройку параметров линейной регрессионной модели выполним минимизируя квадрат суммы отклонения модельных значений прогнозируемой переменной от реальных значений, содержащихся в обучающей выборке, применив при этом L_2 регуляризацию [9, 10].

Функционал ошибки для данного типа регресиионной модели, которую называют гребневой регрессией, имеет вид:

$$Q(\overline{\boldsymbol{w}};(\overline{\boldsymbol{x}}_i,y_i)_{i=1}^{M}) = \sum_{i=1}^{M}(y_i - f(\overline{\boldsymbol{x}}_i))^2 + \lambda\sum_{j=0}^{3} w_j^2 \to \min_{\overline{\boldsymbol{w}}},$$

где $(\overline{\boldsymbol{x}}_i,y_i)_{i=1}^{M}$ – обучающая выборка;

$f(\overline{\boldsymbol{x}}) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$ – функция регрессии;

параметр $\lambda > 0$ – коэффициент регуляризации.

Оценим параметры $\widehat{w}_0$, $\widehat{w}_1, \widehat{w}_2, \widehat{w}_3$ решив уравнение

$$\frac{\partial(Q)}{\partial(w_j)} = 0, j = \overline{0,3}\,. \quad (1)$$

Полученные в результате решения уравнения (1) параметры $\widehat{w}_0$, $\widehat{w}_1, \widehat{w}_2, \widehat{w}_3$ будут некоторыми оценками настоящих w_0, w_1, w_2, w_3 для исследуемой функциональной зависимости. Далее для удобства оценки коэффициентов будем обозначать просто w_0, w_1, w_2, w_3.

Чтобы избежать переобучения разобъем имеющуюся в нашем распоряжении выборку X^n на обучающую и контрольную выборки (X^l, X^k). Контрольную выборку будем использовать только для оценки качества построенной модели. Далее все анализируемые параметры качества модели вычислены по контрольной выборке.

Применим для обучения регрессионной модели метод кросс-валидации по случайным перестановкам, который является базовым инструментом в решении задач машинного обучения. Параметром в данному случае является количество K разбиений выборки на тестовую и обучающую и соотношение количества образцов. Будем использовать разбиение в соотношении 1:3. В качестве результирующей модели выбирается модель с наилучшим значением критерия качества.

Критерием качества модели часто выступает коэффициент детерминации R^2

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

где y_i – реальные значения;

$\widehat{y}_i$ – модельные значения;

$\bar{y}$ – выборочное среднее.

Однако известно, что сравнивать качество модели, опираясь на значение коэффиициента детерминации, в случае, если модели построены по разным выборкам, нельзя. Поэтому в некоторых случаях для оценки качества модели будем использовать среднеквадратичную ошибку и визуальный анализ невязок функции регрессии $\hat{\varepsilon}_i = |y_i - \hat{y}_i|$, а также $E(\hat{\varepsilon})$ и $\sqrt{D(\hat{\varepsilon})}$. Графики для такого визуального анализа будут рассмотены ниже.

Алгоритм поиска аномалий в данных

Выборка, которая находилась в распоряжении авторов, как уже было отмечено ранее, зашумлена, что подтверждено визуальным анализом корреляционных полей. Сырая выборка содержит некоторые нечисловые признаки, которые могут быть использованы экспертом в предметной области для отбраковки некоторых образцов. Чтобы сократить количество образцов, которые должен рассмотреть эксперт, предложен алгоритм поиска аномальных объектов, подозрительных на выборосы. Алгоритм имеет два параметра: K – количество разбиений исходной выборки на тестовую и обучающую, T – допустимая величина отклонения модельного значения от реального, измеряемая в градусах.

Шаг 1. Разбиваем выборку K – раз на обучающую и тестовую в соотношении 3:1:

$$(X_{train}^{l_j}, X_{test}^{l-l_j})_{j=1}^{K}\,. \quad (3)$$

Шаг 2. Строим соответсвенно K регрессионных моделей по подвыборкам $X_{train}^{l_j}$ из (3).

Шаг 3. Для каждой модели $f_j, j = \overline{1,K}$ вычисляем значение $\hat{y}_i = f_j(x_i)$, $i = \overline{1,n}$ и определяем множество индексов образцов Ind_j, таких что

$$\hat{\varepsilon}_i = |y_i - \hat{y}_i| > T, i = \overline{1,n}.$$

Шаг 4. Вычисляем пересечение множеств индексов подозрительных на выбросы объектов каждой модели

$$Ind_{Outliers} = \bigcap_{j=1}^{K} Ind_j$$

В отличие от известных алгоритмов поиска выбросов, таких как IsolationForest или LOF [11], в предложенном алгоритме не нужно задавать в качестве параметра степень загрязненности выбросами. Отметим, что параметр K влияет на количество аномалий, с увеличением K их число уменьшается. Параметр T зависит от свойств данных и задается в соответсвии с требованиями эксперта. В данной задаче T = 8 °C.

Данный алгоритм позволяет определить все выбросы, при условии, что вид восстанавливаемой функциональной зависимости выбран правильно. На рис. 3 изображены результаты работы предложенного алгоритма

для исследуемой выборки. Красным цветом выделены аномалии, подозрительные на выбросы.

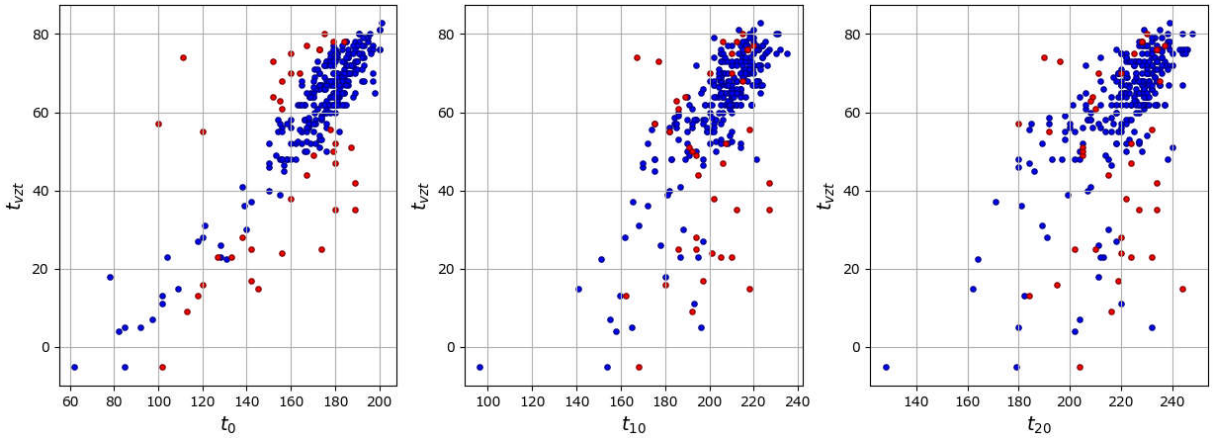

Рисунок 3 – Результат работы метода поиска аномалий

Далее такие аномальные объекты были направлены на исследование экспертам в данной предметной области, чтобы определить, нужно ли их удалять из рассмотрения.

В результате анализа аномальных образцов экспертом была подтверждена необходимость удаления этих объектов из выборки, так как они являлись фальсифицированными или нестандартными продуктами, то есть не являлись дизельным топливом.

Результаты экспериментов

Для решения задачи использовался язык программирования Python 3.5. Регрессионная модель строилась с использованием библиотеки scikit-learn v0.20.

Критерием качества для выбора наилучшей модели является среднеквадратичная ошибка $MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Процедура построения регрессионной модели состояла из трех этапов:

– построение модели на полной выборке, не очищенной от выбросов, количество случайных разбиений $K = 8$;

– поиск аномальных объектов, подозрительных на выбросы;

– построение модели на очищенной выборке, количество случайных разбиений 8.

В табл. 2 приведены результаты первого этапа. Модель с минимально ошибкой имеет вид:

$$f(x) = 0.66x_1 + 0.17x_2 - 0.08x_3 - 70.71. \quad (4)$$

Таблица 2 – Параметры регрессионной модели на полной выборке

Эксперимент (K=8)	***MSE на обучающей выборке***	***MSE на тестовой выборке***	w_0	w_1	w_2	w_3
1	-71.76	-83.11	-77.34	0.62	0.32	-0.16
2	-76.18	-69.64	-79.90	0.62	0.30	-0.12
3	-65.94	-101.56	-75.98	0.67	0.31	-0.19
4	-74.93	-74.04	-72.11	0.66	0.33	-0.22
5	-70.34	-85.79	-71.44	0.64	0.14	-0.02
6	-67.67	-96.71	-71.44	0.63	0.40	-0.27
7	-75.81	**-69.26**	**-70.71**	**0.66**	**0.17**	**-0.08**
8	-75.62	-72.30	-77.17	0.62	0.11	0.05
Среднее значение	-72.28	-81.55	-74.51	0.64	0.26	-0.13
Среднеквадратичное отклонение	3.73	11.65	03.27	3,01	0.10	0.10

Значения коэффициента детерминации и средней квадратичной ошибки на отложенной контрольной выборке X^k:

$$R^2(X^k) = 0.47, MSE(X^k) = 124{,}77. \quad (5)$$

Анализ средних значений и отклонений от среднего для MSE модели на обучении и на тесте по табл. 2 показал нестабильность исходной выборки, что связано с наличием выбросов в ней, а также с ее относительно малым объемом.

На рис. 4 приведены графики невязок построенной модели на обучающих и тестовых данных соответственно. Среднее значение ошибки $E(\hat{\varepsilon}; X^k) = 6.48^{\circ}\text{C}$, среднее отклонение ошибки $D(\hat{\varepsilon}; X^k) = 9.27^{\circ}\text{C}$.

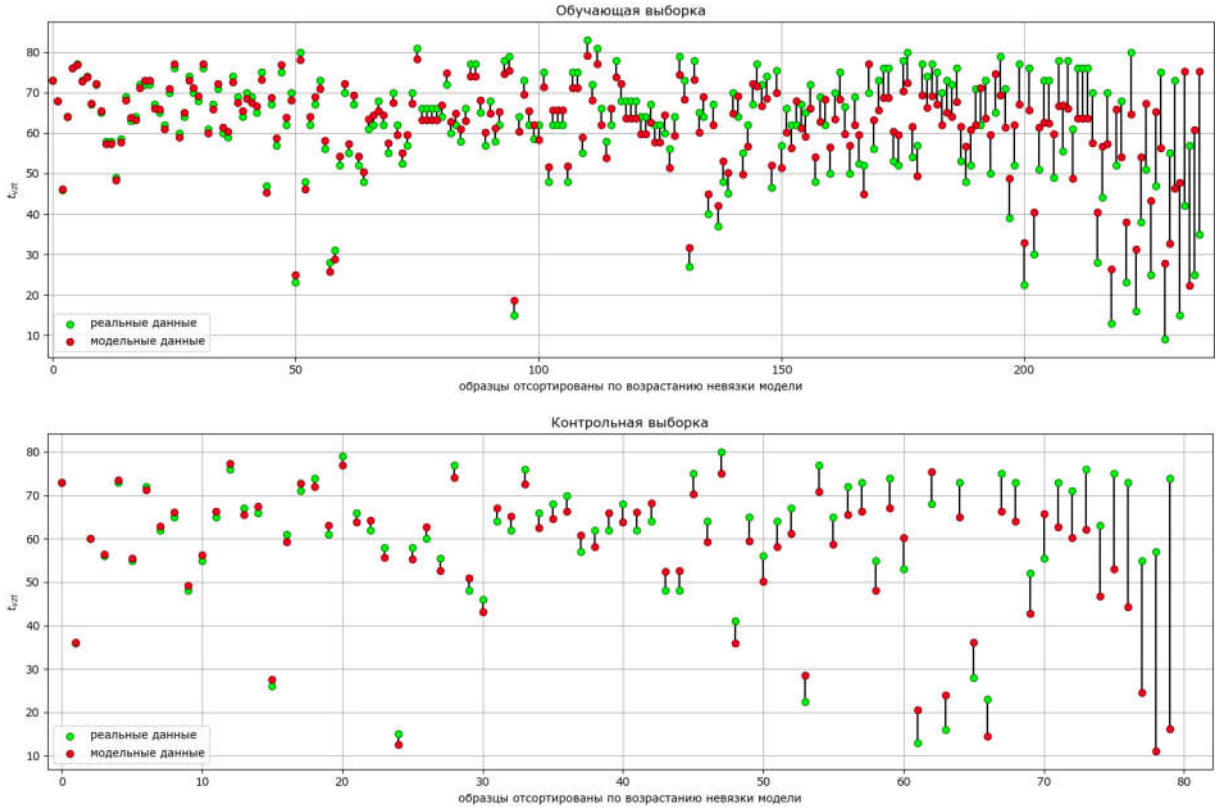

Рисунок 4 – Визуализация качества модели, построенной на полной выборке

Визуальный анализ графиков невязок модели, упорядоченных по возрастанию, позволяет оценить качество модели. Можно видеть, например, что модель ошибается на достаточно небольшом количестве образцов, которые скорее всего являются выбросами.

После выполнения алгоритма поиска аномальных объектов с параметрами $K = 1000$, $T = 8$ отобрано 39 подозрительных на выборосы объектов, которые, после изучения экспертом, были отброшены (рис. 3).

В табл. 3 приведены результаты третьего этапа. Модель, построенная по очищенной от выбросов выборке с минимальной ошибкой имеет вид:

$$f(x) = 0.66x_1 + 0.19x_2 - 0.08x_3 - 74.29. \quad (6)$$

Таблица 3 – Параметры регрессионной модели на очищенной выборке

Эксперимент	***MSE на обучающей выборке***	***MSE на тестовой выборке***	w_0	w_1	w_2	w_3
1	-19.33	-22.41	-69.60	0.64	0.22	-0.11
2	-19.54	-23.14	-81.53	0.67	0.21	-0.07
3	-20.97	**-17.08**	**-74.29**	**0.66**	**0.19**	**-0.08**
4	-20.43	-19.35	-75.18	0.63	0.25	-0.10
5	-20.68	-18.22	-76.10	0.66	0.23	-0.11
6	-20.41	-18.88	-72.58	0.64	0.20	-0.08
7	-19.21	-22.54	-76.38	0.64	0.17	-0.03
8	-18.75	-23.92	-75.27	0.67	0.15	-0.05
Среднее значение	-19.92	-20.69	-75.11	0.65	0.20	-0.08
Среднеквадратичное отклонение	0.75	02.42	03.19	0.02	0.03	0.03

Значения коэффициента детерминации и средней квадратичной ошибки на отложенной контрольной выборке X^k:

$$R^2(X^k) = 0.90, MSE(X^k) = 21{,}25. \qquad (7)$$

Анализ средних значений и отклонений от среднего для MSE модели на обучении и на тесте по табл. 3 показал стабилизацию модели. Значения коэффициентов функции регрессии практически не отличаются от средних коэффициентов w_i. На рис. 5 приведены графики невязок построенной модели на обучающих и тестовых данных соответственно.

Среднее значение ошибки $E(\hat{\varepsilon}; X^k) = 3.83^{\circ}\text{C}$, среднее отклонение ошибки $D(\hat{\varepsilon}; X^k) = 2.68^{\circ}\text{C}$, что удовлетворяет требованиям действующих ГОСТов.

Сравним коэффициенты моделей, полученных по полной и по очищенной выборкам. Значения коэффицинетов (4), (6) приведены в табл. 4. Как можно видеть, коэффициенты w_1 и w_3 совпадают, а w_0 отличается на 3.42 и w_2 на 0.2. Средние значения коэффициентов обеих моделей также очень близки. Это подтверждает тот факт, что построенная модель приближается некоторую неизвестную нам реальную зависимость, восстановление которого является основной задачей машинного обучения.

Таблица 4 – Коэффициенты моделей

	w_0	w_1	w_2	w_3
Коэффициенты лучшей модели, полная выборка	-70.71	0.66	0.17	-0.08
Средние знач. коэффициентов, полная выборка	-74.51	0.64	0.26	-0.13
Коэффициенты лучшей модели, очищенная выборка	-74.29	0.66	0.19	-0.08
Средние значения коэффициентов, очищенная выборка	-75.11	0.65	0.20	-0.08

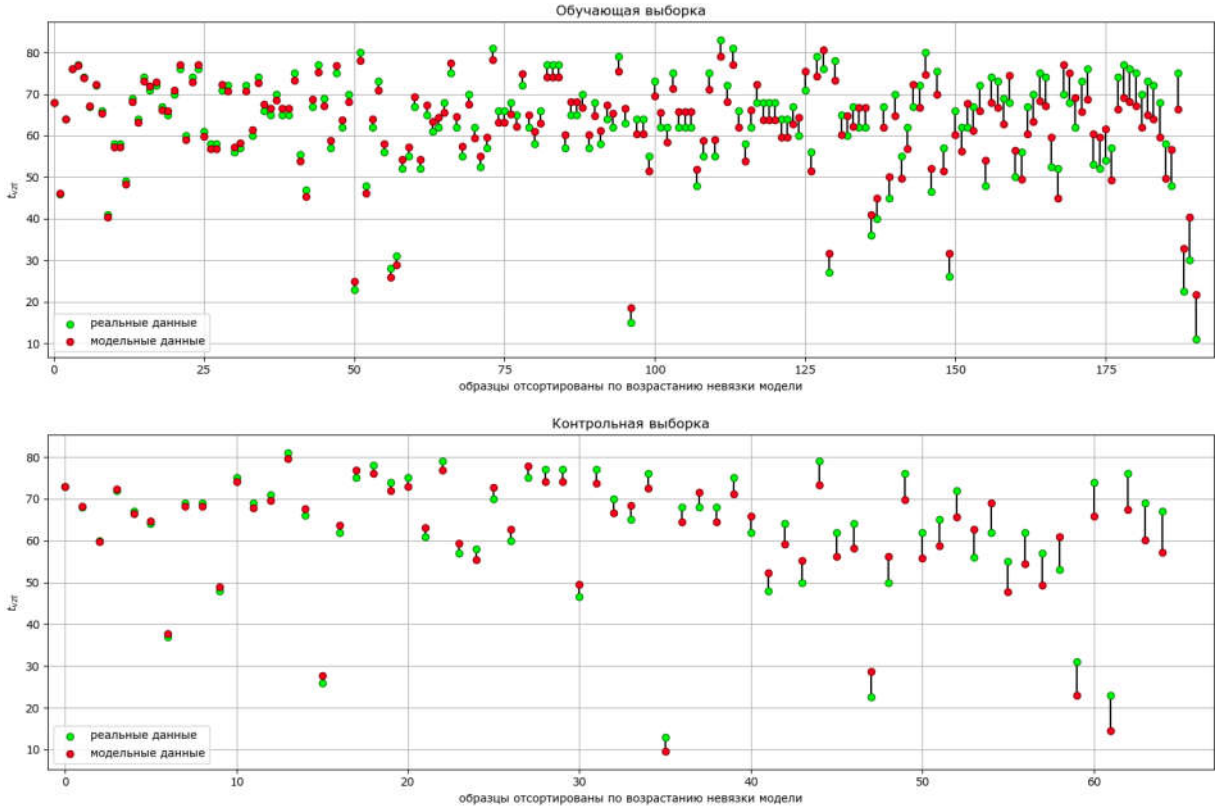

Рисунок 5 – Визуализация качества модели, построенной на очищенной выборке

Вероятностное представление результатов регрессионной модели

Из статистики известно, что коэффициенты модели многомерной линейной регрессии нормально распределены. Когда количество примеров обучающей выборки мало, оценки достоверности регрессионной модели становятся достаточно слабыми. В таких случаях используют байесовский подход для представления коэффициентов модели как нормально распределенных случайных величин $\beta_0, \beta_1, \beta_2, \beta_3$.

Предлагается оценить коэффициенты w_i нормальным распределением. Так как метод кросс-валидации позволяет строить достаточно большое количество моделей, увеличив

количество случайных разбиений обучающей выборки на различные подвыборки до $K = 1000$, посчитаем оценки математических ожиданий и дисперсий данных величин. На рис. 6 представлены гистограммы коэффициентов w_i и восстановленные по статистически оцененным параметрам графики вероятностных распределений случайных величин $\beta_0, \beta_1, \beta_2, \beta_3$.

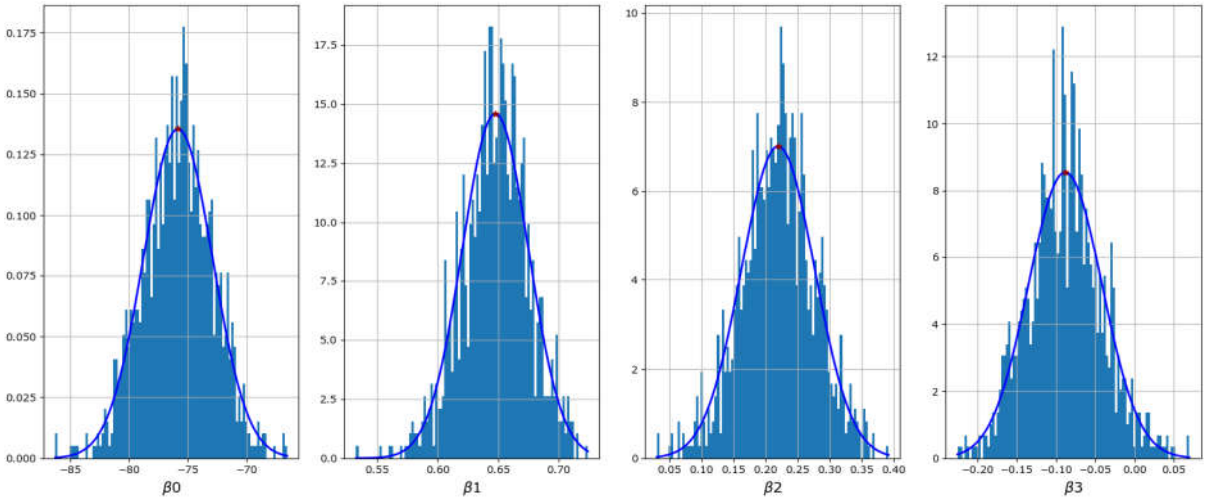

Рисунок 6 – Вероятностные оценки коэффициентов регрессионной модели

Выводы

В данной работе построена многомерная линейная регрессионная модель для вычисления ТВЗТ по температуре кипения 10 % и 20 % фракциям дизельного топлива.

Полученные результаты показали, что тяжелые фракции дизельного топлива не влияют на ТВЗТ, что опровергает результат, полученный в работе [4], где утверждается обратное.

Получить актуальные данные о топливе, которое используется в нашей стране, в настоящее время не представляется возможным. Однако имеющиеся в распоряжении данные за прошлые периоды позволили разработать общий подход к анализу таких данных и процессу построения регрессионной модели. Для усовершенствования модели потребуются новые качественные обучающие данные, которые могут быть получены при затрате определенных ресурсов.

Результат работы может быть использован при анализе похожих данных в области нефтехимии, например, для построения регрессионной модели взаимосвязанных свойств бензинов.

Предложенный метод поиска аномальных объектов показал высокую эффективность при работе с зашумленными данными, содержащими грубые ошибки, обусловленные, в том числе, человеческим фактором. Качество построенной модели полностью удовлетворяет требованиям существующих ГОСТов.

Так как выборка была сильно зашумлена, качество модели несколько хуже, чем в работах, проанализированных бразильскими авторами в [4]. В этой работе использованы чистые образцы, полученные непосредственно на нефтеперерабатывающих заводах. В нашем случае даже привезенные в лабораторию «чистые» образцы были подвержены смешению с остатками в цистернах для транспортировки или в емкостях хранения. В целом контроль качества нефтепродуктов в нашей стране подвержен ослаблению, что вызвано нежеланием со стороны потенциальных заказчиков проводить мероприятия такого направления, но, с другой стороны, подтверждает актуальность исследований призванных повысить его эффективность, так как топлива высокого качества необходимы для стабильного развития экономики.

Литература

1. ГОСТ 32511-2013 Топливо дизельное ЕВРО. Технические условия [Текст]. Введ. 2015─01─01.─ М.: Стандартинформ, 2014. - 20 с.

2. Пучков Н. Г. Дизельные топлива. – М.: Государственное научно-техническое издательство нефтяной и горно-топливной литературы. – 1953. – 194 с.

3. Arankalle A. Significance of flash point in diesel fuel specification. 3rd, International conference on synergy of fuel and automotive technology for a cleaner environment // SAE 2004 India Mobility Conference; India in International conference on synergy of fuel and automotive technology for a cleaner environment. – New Delhi: Allied Publishers. – 2004. – PP. 508-512.

4. Aleme H. G., Barberi P. J. S. Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration // Fuel. – 2012. – V. 102. – PP. 129–134.

5. Александрова И. А. Перегонка и ректификация в нефтепереработке. – М. Химия, 1981 г. – 352 с., ил.

6. Ishida H., Iwarna A. Some Critical Discussions On Flash And Fire Points Of Liquid Fuels // Fire Safety Science 1, 1986: – P. 217-226. doi:10.3801/IAFSS.FSS.1-217

7. Laurent Catoire A Unique Equation to Estimate Flash Points of Selected Pure Liquids Application to the Correction of Probably Erroneous Flash Point Values // Journal of Physical and Chemical Reference Data, Vol. 33, №4, December, 2004. – P. 1083-1112.

8. ДСТУ 3868-99 «Топливо дизельное. Технические условия», 1993.

9. Айвазян С. А., Мхитарян В. С. Теория вероятностей и прикладная статистика. – М.: ЮНИТИ-ДАНА, 2001. – Т.1. – 656 с.

10 Айвазян С. А. Теория вероятностей и прикладная статистика. – Т.2 – М.: ЮНИТИ-ДАНА, 2001. – 432 с.

11. Breunig M. M., Kriegel H. P., Ng R. T., Sander J. LOF: identifying density-based local outliers // ACM sigmod record, 2000. – P. 1-12.

Максимова А. Ю., Иванова А. А., Лозинский Н. С. Регрессионная модель для прогнозирования температуры вспышки дизельного топлива в закрытом тигле. *Рассматривается задача построения функциональной зависимости температуры вспышки дизельного топлива в закрытом тигле от его измеренных нормативных параметров с использованием подходов машинного обучения. Выполнен анализ сырых данных лаборатории контроля качества методами корреляционного анализа. Разработан метод поиска аномальных объектов, подозрительных на выбросы. Применена гребневая регрессионная модель, построены вероятностные оценки параметров линейной модели. Предложенный подход может быть применен при анализе похожих данных в области нефтехимии.*

Ключевые слова: *регрессионная модель, машинное обучение, температура вспышки в закрытом тигле.*

Maksimova A., Ivanova A., Losinskyy N. A regression model for predicting the flash point of diesel in a closed crucible. *The problem of constructing a functional dependence of the flash point of diesel fuel in a closed crucible on its measured normative parameters using machine learning approaches is considered. The analysis of raw data of the laboratory of quality control by the methods of correlation analysis is performed. A method has been developed to search for anomalous objects suspicious of outliers. A ridge regression model is applied, and probabilistic estimates of the parameters of the linear model are constructed. The proposed approach can be applied in the analysis of similar data in the field of petrochemicals.*

Keywords: *regression model, machine learning, flash point of diesel in a closed crucible.*

Статья поступила в редакцию 11.12.2019
Рекомендована к публикации профессором Миненко А. С.