УДК 004.4

О. Экономическое развитие, инновации, технологические изменения и рост
O. Economic Development, Innovation, Technological Change, and Growth

ПРИМЕНЕНИЕ МНОГОПАРАМЕТРИЧЕСКИХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОТБОРА ЗНАЧИМЫХ КОЛИЧЕСТВЕННЫХ ХАРАКТЕРИСТИК ПРЕДПРИЯТИЙ В РЕГИОНАХ РОССИЙСКОЙ ФЕДЕРАЦИИ ПРИ АНАЛИЗЕ ДОТАЦИОННОСТИ

А. В. КУЗНЕЦОВА[1)]***, Л. Р. БОРИСОВА***[2)]

[1)]*Институт биохимической физики им. Н. М. Эмануэля РАН, ул. Косыгина, 4, 119334, г. Москва, Россия*
[2)]*Финансовый университет при Правительстве Российской Федерации, пр. Ленинградский, 49, 125167, г. Москва, Россия*

Аннотация. Представлен оригинальный метод поиска связи финансово-экономических показателей с дотационностью регионов Российской Федерации. Как наиболее значимые с точки зрения потребностей регионов в дотациях параметры рассмотрены данные по предприятиям и организациям, а также касающиеся основных фондов показатели. Выделены две группы регионов: регионы, нуждающиеся в дотациях, и регионы, не нуждающиеся в дотациях. Методами машинного обучения в выделенных группах выявлены различия по отчетным данным предприятий и организаций, а также основным фондам. Наиболее важными показателями за 2020 г., по которым группы отличаются друг от друга, стали число и оборот организаций, сальдированный финансовый результат (разность прибыли и убытка), удельный вес убыточных организаций, кредиторская и дебиторская задолженности организаций, просроченная задолженность по заработной плате в расчете на одного работника, число малых предприятий на 10 000 человек населения и др. Такой подход (классификация методами оптимально достоверных разбиений и статистически взвешенных синдромов) только начинает использоваться в данной области. Найденные закономерности позволят более точно обрисовать паттерн («портрет») каждого региона Российской Федерации и дадут возможность прогнозировать их дотационный статус в будущем. Набор значимых характеристик позволит повысить точность прогноза и разработать план по выходу из группы регионов, нуждающихся в дотациях, в группу регионов, не нуждающихся в дотациях.

Ключевые слова: методы машинного обучения; статистика; *data science*; экономические показатели; дотационные регионы; субъекты Российской Федерации.

Образец цитирования:
Кузнецова АВ, Борисова ЛР. Применение многопараметрических методов машинного обучения для отбора значимых количественных характеристик предприятий в регионах Российской Федерации при анализе дотационности. *Журнал Белорусского государственного университета. Экономика*. 2025;1:88–96.
EDN: JITKVA

For citation:
Kuznetsova AV, Borisova LR. Application of multiparametric machine learning methods for selection of significant quantitative characteristics of enterprises in the regions of the Russian Federation in the analysis of subsidisation. *Journal of the Belarusian State University. Economics.* 2025;1:88–96. Russian.
EDN: JITKVA

Авторы:
Анна Викторовна Кузнецова – кандидат биологических наук; старший научный сотрудник лаборатории математической биофизики.
Людмила Робертовна Борисова – кандидат физико-математических наук, доцент; доцент кафедры математики и анализа данных факультета информационных технологий и анализа больших данных.

Authors:
Anna V. Kuznetsova, PhD (biology); senior researcher at the laboratory of mathematical biophysics.
azforus@yandex.ru
https://orcid.org/0000-0002-0297-7013
Ludmila R. Borisova, PhD (physics and mathematics), docent; associate professor at the department of mathematics and data analysis, faculty of information technology and big data analysis.
lrborisova@fa.ru
https://orcid.org/0000-0002-5757-0341

APPLICATION OF MULTIPARAMETRIC MACHINE LEARNING METHODS FOR SELECTION OF SIGNIFICANT QUANTITATIVE CHARACTERISTICS OF ENTERPRISES IN THE REGIONS OF THE RUSSIAN FEDERATION IN THE ANALYSIS OF SUBSIDISATION

***A. V. KUZNETSOVA*[a], *L. R. BORISOVA*[b]**

[a]*Emanuel Institute of Biochemical Physics of Russian Academy of Sciences, 4 Kosygina Street, Moscow 119334, Russia*
[b]*Financial University under the Government of the Russian Federation, 49 Leningradskij Avenue, Moscow 125167, Russia*
Corresponding author: L. R. Borisova (lrborisova@fa.ru)

Abstract. In this paper, we present an original method for searching a connection between financial-economic characteristics and the subsidisation of the regions of the Russian Federation. The dataset contained data on enterprises and organisations, as well as indicators related to fixed assets, as the most significant in terms of the regions' needs for subsidies. Two groups of regions were identified: regions with high subsidies and regions without them. Machine learning methods were used to establish differences in the reporting data of enterprises and organisations, as well as fixed assets, in the identified groups. In 2020, the most important indicators by which the groups differed from each other were the number and turnover of organisations, the balanced financial result (difference of profit and loss), the share of unprofitable organisations, accounts payable and receivable of organisations, overdue wage arrears per employee, the number of small enterprises per 10 000 people, etc. This approach (classification using optimally reliable partitioning and statistically weighted syndromes) is just beginning to be used in this area. The found dependences will allow us to more accurately outline the pattern («portrait») of each region of the Russian Federation with the possibility of further forecasting its subsidised status. A set of significant characteristics will improve the accuracy of the forecast and propose a plan for moving from the subsidised group to the group of self-sufficient subjects of the Russian Federation.

Keywords: machine learning methods; statistics; data science; economic indicators; subsidised regions; subjects of the Russian Federation.

Введение

В Российскую Федерацию входят 89 регионов, которые находятся в разных климатических поясах, отличаются по степени освоенности, количеству полезных ископаемых, плотности населения. Актуальным вопросом становится анализ дотационных регионов, выявление основных критериев, по которым отличаются субъекты Российской Федерации с разной степенью экономической самостоятельности.

В работе [1] приводится наиболее полная с точки зрения негативного влияния на бюджетное состояние и уровень дотационности регионального бюджета классификация социально-экономических факторов. Нивелировать проблемы региональной диспропорции и социально-экономической дифференциации мезоуровня возможно лишь с учетом множества детерминант, которые субъективно и (или) объективно систематически воздействуют на текущее и перспективное состояние регионов страны, изменяют их бюджетные характеристики. Для анализа причин регионального неравенства важно выявить и сгруппировать регионы, наиболее зависимые от федерального центра. Для перехода современной России из категории ведущих развивающихся стран (БРИКС) в категорию развитых стран (ОЭСР) необходимо обеспечить сбалансированное состояние региональной экономики.

В Российской Федерации регионов-доноров в три раза меньше, чем дотационных регионов. На них, кроме обеспечения собственных трат, ложится нагрузка по финансированию дотационных субъектов (у регионов-доноров доход, а соответственно, и перечисления в российский бюджет превышают их потребности и расходы). В регионах-донорах нет проблем с инвестициями, часто в них находятся крупные залежи полезных ископаемых, активно развивается промышленность, изготавливается продукция с высокой прибавочной стоимостью, идущая на экспорт. В первую пятерку регионов-доноров входят два региона с крупными залежами полезных ископаемых и два мегаполиса с высокой инвестиционной привлекательностью, развитой финансовой сферой, технологичными производствами. В порядке убывания поступлений в бюджет из этих регионов их список выглядит следующим образом: Ханты-Мансийский автономный округ, Москва, Ямало-Ненецкий автономный округ, Санкт-Петербург, Московская область. К ним можно добавить такие хорошо развитые регионы, как Свердловская область, Татарстан, Красноярский край, а также Липецкую и Кемеровскую области.

У основной части регионов Российской Федерации уровень дотационной поддержки составляет примерно 19–20 %. При этом существуют хронически дотационные регионы, большая часть которых в течение длительного времени формируют свой бюджет за счет дотаций из федерального центра более чем на 50 %. В той или иной степени помощи из центра требуют 68 регионов. Основной причиной такого положения является отсутствие у чиновников мотивации предпринимать какие-то действия, чтобы изменить сложившуюся ситуацию, когда из центра стабильно поступает необходимое количество денежных средств. Немалое число регионов даже при наличии значительных поступлений из федерального бюджета и огромных собственных ресурсов находятся в упадке и не развиваются [2; 3].

Обозначенную проблему можно решить путем проведения комплексных мероприятий для развития данных регионов: стимулирования инвестиций в экономику, развития инфраструктуры, обучения и повышения квалификации населения, поддержки развития сельского хозяйства и т. д. Внедрение таких мер позволит снизить зависимость дотационных регионов от федерального бюджета, стимулировать их экономическое развитие и улучшить качество жизни населения.

Необходимо создать механизм оценки дотационности регионов Российской Федерации и выявить основные слабые стороны, для чего можно применить современные подходы и методы *data science*, основанные на многопараметрическом анализе данных. *Data science* играет ключевую роль в создании механизма оценки дотационности регионов.

Методология машинного обучения хорошо изложена в работе [4] на примере оценки инвестиционной деятельности различных регионов Российской Федерации с использованием программного продукта *MatLab 2018b*. В ней сделан акцент на том, что алгоритм обучения применяется для обнаружения в данных знаний или свойств и их изучения. Качество или количество данных будут влиять на эффективность обучения и прогнозирования. Машинное обучение представляет собой метод, который подразумевает работу с данными. Эта система меняется вместе с изменением входящих в нее данных. В идеале она должна одновременно подсказывать пути изменения характеристик объектов, которые позволят им отвечать нужным требованиям. Такой подход принято называть *data science*, а специалистов этого профиля – *data scientists*. Методы машинного обучения используются в различных областях, например в менеджменте при оценке риска, производстве, разработке рекомендательных систем [5–7].

В данной работе мы предлагаем как результаты применения всем известных методов машинного обучения, работающих в режиме автоклассификации системы анализа данных *data master azforus*, так и оригинальные методы, обладающие определенными преимуществами [8–10]. Одним из таких преимуществ является прозрачное решение, позволяющее наглядно представлять объекты на диаграммах рассеяния в окружении похожих объектов, выявлять набор значимых показателей и рекомендовать конкретные действия по управлению объектами в нужном направлении.

Методами машинного обучения на обучающей выборке был выделен набор наиболее информативных показателей, так или иначе влияющих на отнесение предприятия к определенной группе: группе регионов, нуждающихся в дотациях (группа 1), или группе регионов, не нуждающихся в дотациях (группа 2). В ходе исследований использовались различные методы машинного обучения, в том числе оригинальные логико-статистические методы (*data science*) [11], оценивалась достоверность найденных закономерностей, основанных на построении оптимальных достоверных разбиений признакового пространства [12]. Использование логико-статистических методов позволило провести анализ, не делая априорных предположений о виде вероятностных распределений. Эти методы также эффективны при работе с выборкой любого размера и большим количеством плохо структурированных признаков.

Материалы и методы исследования

В базу данных для исследования вопросов дотационности вошли 46 регионов, разделенных на две группы. Для проведения контроля решающего правила оставили 36 регионов. В группу 1 вошли 22 дотационных региона: Дагестан, Ингушетия, Кабардино-Балкария, Карачаево-Черкесия, Северная Осетия (Алания), Чечня, Башкортостан, Тыва, Бурятия, Якутия, Крым, Чувашия, Ставропольский, Забайкальский, Алтайский, Камчатский края, Брянская, Ивановская, Ростовская, Кировская, Курганская области, Чукотский автономный округ.

В группу 2 вошли 24 региона, не нуждающихся в дотациях: города Москва и Санкт-Петербург, Белгородская, Калужская, Липецкая, Московская, Тульская, Ярославская, Вологодская, Ленинградская, Мурманская, Нижегородская, Самарская, Свердловская, Пермская, Сахалинская, Тюменская, Иркутская области, Ненецкий, Ханты-Мансийский, Ямало-Ненецкий автономные округа, Краснодарский и Красноярский края, Татарстан.

Финансово-экономические показатели были отобраны на сайте Федеральной службы государственной статистики из баз данных, представленных в ежегодных сборниках «Регионы России» за 2020 и 2021 гг. Количество финансово-экономических показателей составило 14. Валидация осуществлялась методом скользящего контроля (*leave-one-out*). Анализу подверглись 47 регионов, 22 из них отнесены к группе 1, 25 – к группе 2 (табл. 1).

При анализе данных методами машинного обучения были использованы шесть традиционных методов, входящих в режим автоклассификации системы анализа данных *data master azforus*: адаптивный бустинг, деревья решений, градиентный бустинг, метод ближайших соседей, метод опорных векторов, линейный дискриминантный анализ. При использовании режима автоклассификации методы ранжируются по результатам ROC AUC. Далее на лучших наборах методов вычисляется ансамбль методов, т. е. их совокупность. Сравниваются результаты ансамблей из трех, пяти, семи или девяти методов. Лучший ансамбль отбирается для решающего правила.

Кроме перечисленных методов машинного обучения, были использованы также два оригинальных метода: метод оптимально достоверных разбиений (метод ОДР) и метод статистически взвешенных синдромов (метод СВС). Метод ОДР основан на разбиении пространства значений показателя границей разбиения таким образом, чтобы с одной стороны от нее преобладали объекты одной группы сравнения, а с другой стороны – объекты другой группы. Значимость найденной закономерности оценивалась с помощью перестановочного теста Монте-Карло. Этот метод очень затратен по времени, если число объектов больше тысячи, но для задач с небольшим числом объектов (как в случае с регионами) он обладает неоспоримым преимуществом, поскольку решает проблему черного ящика. Из всего объема данных отбираются только значимые показатели. Классификация проводится методом СВС, осуществляющим голосование по базовым множествам. Автономная программа «Прогноз» на основе решающего правила метода СВС позволяет вводить данные одного или нескольких объектов вручную, загружать их файлом или брать из обучающей выборки. После распознавания на наглядных диаграммах рассеяния можно наблюдать расположение объекта в кругу похожих на него объектов. Также можно создать план перевода объекта из группы 1 (регион, требующий дотаций) в группу 2 (регион, обеспечивающий себя самостоятельно).

При использовании любого метода распознавание произвольного объекта может быть представлено в виде последовательного выполнения двух операций. На первом шаге вычисляются так называемые оценки за группы (классы) – вещественные величины, отражающие меру родства объекта классу. На втором шаге производится собственно распознавание. Обычно при этом используется пороговое правило: если оценка объекта за класс K больше некоторого порога d, то объект относится к классу K, если оценка объекта за класс K меньше порога d, то объект не относится к классу K. Для многих методов оценки обычно вычисляются таким образом, что в результате они принадлежат отрезку [0; 1]. В этом случае оценки могут интерпретироваться как вероятности принадлежности классам. Обычно в задачах распознавания с двумя непересекающимися классами сумма оценок за эти классы равна 1. Все показатели эффективности, кроме ROC AUC, зависят от величины порога d. Как правило, наилучшим и наиболее сбалансированным является коллективное решение с максимальным значением по показателю *accuracy* и F-критерию, а также с высокими значениями по другим показателям.

Результаты и их обсуждение

Применение традиционных методов статистики. Использование теста Манна – Уитни – Уилкоксона свидетельствует, что 11 из 14 показателей заметно различаются для групп сравнения. Для каждой группы приведены средние значения (см. табл. 1, значение 1 и значение 2). Три показателя (удельный вес убыточных организаций, просроченная задолженность по заработной плате в расчете на одного работника, число индивидуальных предпринимателей на 10 000 человек населения) оказались незначимыми.

Наиболее важными показателями стали оборот организаций, сальдированный финансовый результат (разность прибыли и убытка), дебиторская и кредиторская задолженности организаций, число малых предприятий на 10 000 человек населения, среднесписочная численность работников, выручка от реализации товаров (работ, услуг) малых предприятий. По всем этим показателям дотационные регионы имеют более низкие значения, чем регионы, которые сами себя обеспечивают. Приведенное в последнем столбце таблицы p-значение близко к нулю для всех исследуемых показателей. Этот результат свидетельствует, что две группы существенно отличаются друг от друга по этим показателям.

Таблица 1

Результаты теста Манна – Уитни – Уилкоксона

Table 1

Mann – Whitney – Wilcoxon test results

Показатели	Количество регионов группы 1	Значение 1	Количество регионов группы 2	Значение 2	*p*-Значение
Число организаций	22	22 289,5	25	84 687,6	0,001
Оборот организаций, млрд руб.	22	480,324	25	5510,4	0
Сальдированный финансовый результат, млн руб.	22	33 055,4	25	485 542,8	0
Кредиторская задолженность организаций, млн руб.	22	182 284,7	25	2 170 680,6	0
Дебиторская задолженность организаций, млн руб.	22	125 035,5	25	2 175 510,0	0
Число малых предприятий на 10 000 человек населения	22	81,316	25	141,6	0
Среднесписочная численность работников	22	56,736	25	238,6	0
Выручка от реализации товаров (работ, услуг) малых предприятий, млн руб.	22	267,664	25	1501,1	0
Численность работников, занятых в бизнесе индивидуального предпринимателя, в 2021 г.	22	59,482	25	140,1	0,003
Средняя численность работников, занятых в бизнесе индивидуального предпринимателя	22	50,895	25	101,9	0,006
Выручка от реализации товаров в индивидуальном предпринимательстве, млн руб.	22	127,168	25	327,4	0,003

Примечание. Здесь и в табл. 2, 3 приводятся показатели за 2020 г., если не указано иное.

Одномерные разбиения. Метод ОДР применялся для показателей, указанных в табл. 1. Был выявлен такой же набор значимых показателей. Преимуществом данного метода является получение границ разбиения, которые отделяют значения одной из исследуемых групп от значений другой группы. В табл. 2 названные границы приведены во втором столбце. Знания о них позволяют наметить план вывода объекта из группы 1 в группу 2. Отметим, что приведенные в предпоследнем столбце табл. 2 значения *F*-критерия влияют на разделение групп: чем больше значение *F*-критерия, тем более очевидно разделение групп.

Таблица 2

Одномерные разбиения методом оптимально достоверных разбиений

Table 2

One-dimensional partitions by the method of optimally reliable partitions

Показатели	Граница разбиения	Квадрант 1		Квадрант 2		*F*-критерий	*p*-Значение
		Количество (доля) регионов группы 1	Количество (доля) регионов группы 2	Количество (доля) регионов группы 1	Количество (доля) регионов группы 2		
Оборот организаций, млн руб.	947,4	20 (90,9 %)	2 (8 %)	2 (9,1 %)	23 (92 %)	31,620	0,00050
Общая дебиторская задолженность организаций, млн руб.	135 110	18 (81,8 %)	1 (4 %)	4 (18,2 %)	24 (96 %)	28,800	0,00050
Общая кредиторская задолженность организаций, млн руб.	198 461	18 (81,8 %)	1 (4 %)	4 (18,2 %)	24 (96 %)	28,800	0,00050
Сальдированный финансовый результат, млн руб.	61 624	19 (86,4 %)	5 (20 %)	3 (13,6 %)	20 (80 %)	20,190	0,00050
Выручка от реализации товаров (работ, услуг) малых предприятий, млн руб.	308,3	17 (77,3 %)	4 (16 %)	5 (22,7 %)	21 (84 %)	17,400	0,00050
Число малых предприятий на 10 000 человек населения	113,5	19 (86,4 %)	7 (28 %)	3 (13,6 %)	18 (72 %)	15,780	0,00067

Окончание табл. 2
Ending of the table 2

Показатели	Граница разбиения	Квадрант 1		Квадрант 2		*F*-критерий	*p*-Значение
		Количество (доля) регионов группы 1	Количество (доля) регионов группы 2	Количество (доля) регионов группы 1	Количество (доля) регионов группы 2		
Среднесписочная численность работников	91,25	18 (81,8 %)	7 (28 %)	4 (18,2 %)	18 (72 %)	13,320	0,001 00
Численность работников, занятых в бизнесе индивидуального предпринимателя, в 2021 г.	60,8	17 (77,3 %)	6 (24 %)	5 (22,7 %)	19 (76 %)	13,010	0,002 00
Число организаций	29 985	17 (77,3 %)	7 (28 %)	5 (22,7 %)	18 (72 %)	11,130	0,007 67
Выручка от реализации товаров в индивидуальном предпринимательстве, млн руб.	164,2	17 (77,3 %)	8 (32 %)	5 (22,7 %)	17 (68 %)	9,428	0,020 33
Средняя численность работников, занятых в бизнесе индивидуального предпринимателя	51,5	16 (72,7 %)	7 (28 %)	6 (27,3 %)	18 (72 %)	9,169	0,028 00

Оборот организаций, дебиторская и кредиторская задолженности организаций, сальдированный финансовый результат, выручка от реализации товаров – наиболее значимые показатели из всей исследуемой базы.

Двумерные разбиения. Двумерные разбиения позволяют более детально выявить значимые закономерности на парах показателей. К рассматриваемым значимым показателям добавились еще удельный вес убыточных организаций и число индивидуальных предпринимателей на 10 000 человек населения (табл. 3, рис. 1). Выявлена отрицательная корреляция показателей для группы 2 (см. рис. 1).

Таблица 3

Двумерные разбиения

Table 3

Two-dimensional partitions

Номер пары показателей	Показатели	Хи-квадрат	Граница разбиения	*p*-Значение
1	Число организаций	23,55	29 985	0,022 670
	Удельный вес убыточных организаций, %	–	38,6	0,000 333
2	Число организаций	21,32	30 536	0,045 000
	Число индивидуальных предпринимателей на 10 000 человек населения	–	202	0,001 000
3	Сальдированный финансовый результат, млн руб.	32,52	–14 128	0,000 500
	Общая кредиторская задолженность организаций, млн руб.	–	182 128	0,029 000
4	Удельный вес убыточных организаций, %	23,71	39,3	0,000 670
	Среднесписочная численность работников	–	80,7	0,048 670
5	Удельный вес убыточных организаций, %	27,55	39,85	0,000 500
	Выручка от реализации товаров (работ, услуг) малых предприятий, млн руб.	–	308,3	0,033 670
6	Удельный вес убыточных организаций, %	23,71	39,3	0,000 500
	Численность работников, занятых в бизнесе индивидуального предпринимателя, в 2021 г.	–	60,8	0,047 670
7	Удельный вес убыточных организаций, %	20,26	39,3	0,002 330
	Средняя численность работников, занятых в бизнесе индивидуального предпринимателя	–	38,7	0,015 000

Окончание табл. 3
Ending of the table 3

Номер пары показателей	Показатели	Хи-квадрат	Граница разбиения	*p*-Значение
8	Общая дебиторская задолженность организаций, млн руб.	34,91	152 910	0,027 670
	Число индивидуальных предпринимателей на 10 000 человек населения	–	230,5	0,000 500
9	Общая кредиторская задолженность организаций, млн руб.	31,94	182 128	0,028 670
	Общая дебиторская задолженность организаций, млн руб.	–	73 444	0,001 000

Примечание. Прочерк обозначает, что вычисления по методу «хи-квадрат» для безразмерных признаков не проводились.

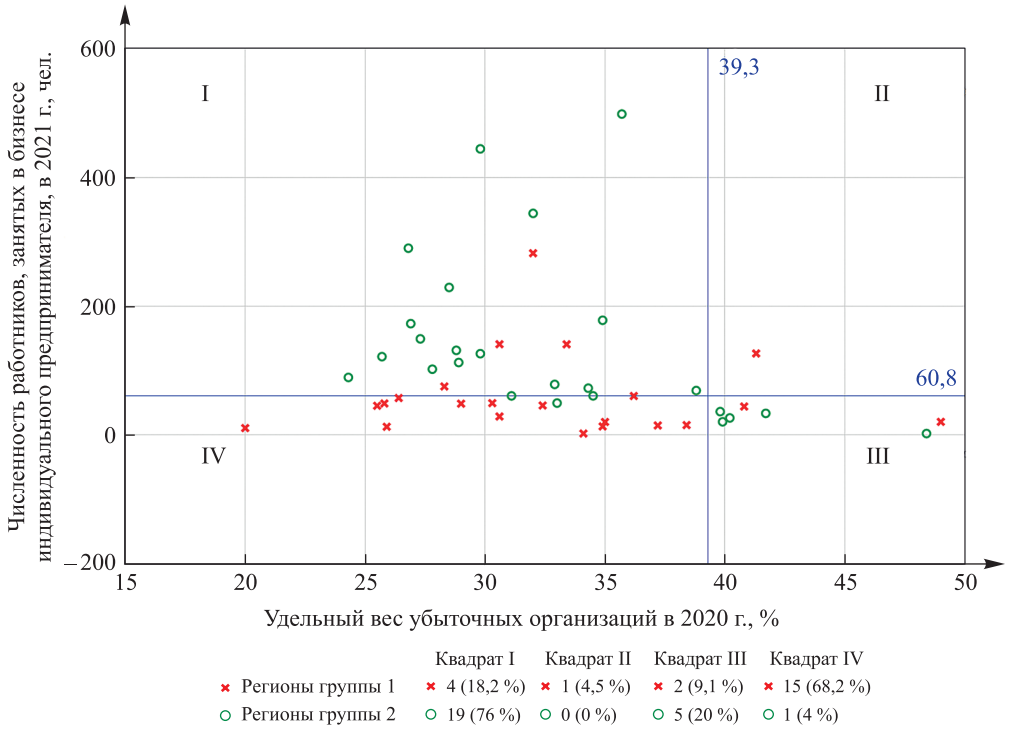

Рис. 1. Двумерная диаграмма рассеяния (*p*-значение по оси абсцисс равно 0,047 666 67, *p*-значение по оси ординат составляет менее 0,000 5)

Fig. 1. Two-dimensional hattering diagram (*p*-value on axis abscissa is 0.047 666 67, *p*-value on axis ordinate is less than 0.000 5)

Результаты автоклассификации. Результаты использования методов машинного обучения, участвовавших в автоклассификации, расположены по убывающей. Самыми эффективными оказались методы, указанные в табл. 4. Наилучшее значение основного показателя качества распознавания объектов при скользящем контроле ROC AUC составляет 0,915 (чем ближе этот показатель к единице, тем лучше распознавание).

Таблица 4

Результаты распознавания семи методов машинного обучения на скользящем контроле

Table 4

Recognision results of seven machine learning methods based on sliding control

Метод	Чувствительность	Специфичность	*F*-оценка	AUC
Адаптивный бустинг	0,909	0,920	0,909	0,915
Метод статистически взвешенных синдромов	0,818	0,800	0,800	0,877
Деревья решений	0,864	0,880	0,864	0,872
Градиентный бустинг	0,773	0,920	0,829	0,869
Метод ближайших соседей	0,773	0,920	0,829	0,846
Метод опорных векторов	0,773	0,760	0,756	0,835
Линейный дискриминантный анализ	0,818	0,640	0,735	0,729

Заметим, что *F*-оценка – мера точности работы модели, основанная на метриках *precision* и *recall*. Чем ближе значение *F*-оценки к единице (100 %), тем лучше модель справляется с задачей классификации. Если *F*-оценка равна нулю, модель полностью не справляется с задачей классификации. Этот показатель отличается от *F*-критерия (см. табл. 2).

Результаты исследования свидетельствуют, что Башкортостан, Ставропольский край и Ростовская область имеют достаточный потенциал, чтобы выйти из группы 1. Ярославская область и Ненецкий автономный округ, хотя и находятся в группе 2, рискуют потерять свой статус (табл. 5, рис. 2).

Таблица 5

Результаты применения ансамбля из пяти методов

Table 5

The results of the ensemble of five methods

Группа	Общее количество (доля) объектов	Количество (доля) правильно включенных объектов	Количество (доля) ошибочно включенных объектов
Группа 1	22 (46,8 %)	19 (86,4 %)	3 (13,6 %)
Группа 2	25 (53,2 %)	23 (92,0 %)	2 (8,0 %)
Всего	*47 (100 %)*	*42 (89,4 %)*	*5 (10,6 %)*

Чистый контроль. Группу регионов, которые в начале исследования не вошли в обучающую выборку, подвергли распознаванию с помощью автономной программы «Прогноз» с «зашитым» в нее решающим правилом.

Из 34 субъектов Российской Федерации 25 были отнесены к группе 1 (Архангельская область, Архангельская область без Ненецкого автономного округа, Орловская, Рязанская, Смоленская, Тамбовская, Тверская, Новгородская, Псковская, Пензенская, Саратовская, Ульяновская, Астраханская, Томская, Магаданская области, Тюменская область без автономных округов, Еврейская автономная область, г. Севастополь, республики Марий Эл, Мордовия, Карелия, Адыгея, Калмыкия, Хакасия, Коми).

К группе 2 алгоритмом распознавания были отнесены Калининградская, Новосибирская, Челябинская, Волгоградская, Омская, Амурская, Удмуртская области, Приморский и Хабаровский края.

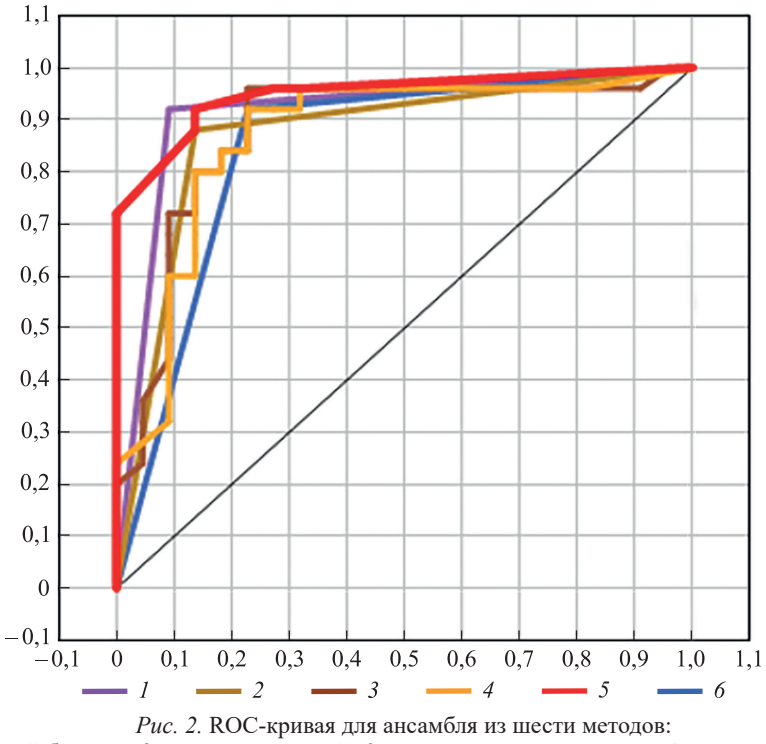

Рис. 2. ROC-кривая для ансамбля из шести методов:
1 – адаптивный бустинг; *2* – деревья решений; *3* – метод опорных векторов; *4* – градиентный бустинг; *5* – линейный дискриминантный анализ; *6* – метод ближайших соседей

Fig. 2. ROC-curve for an ensemble of six methods:
1 – adaptive boosting; *2* – decision trees; *3* – support vector machine; *4* – extreme gradient boosting; *5* – linear discriminant analysis; *6* – *k*-nearest neighbours algorithm

Заключение

Результаты анализа, проведенного методами машинного обучения, свидетельствуют, что наиболее значимыми показателями для исследуемых регионов стали число и оборот организаций, сальдированный финансовый результат, удельный вес убыточных организаций, дебиторская и кредиторская задолженности организаций, просроченная задолженность по заработной плате в расчете на одного работника, число малых предприятий на 10 000 человек населения. Авторы статьи считают, что, поскольку предложенный математический подход показал эффективность в поиске ключевых социально-экономических факторов, можно применить аналогичную технологию для анализа более полной базы данных, содержащей информацию о промышленных и других предприятиях, образовательных учреждениях, социальных службах и объектах, привлекательных для инвестиций.

Метод ОДР и метод СВС позволяют создать план перевода объекта из группы 1 в группу 2 и показывают, на какие ключевые показатели необходимо обращать внимание в первую очередь и каких границ требуется достичь для того, чтобы увеличить вероятность благоприятного прогноза.

Для каждого региона группы 1 можно создать четкий план по переводу его в группу 2. В данный план должны входить 20 значимых показателей с границами разбиения, которые нужно перейти. После достижения этой цели вероятность отказа региона от дотаций из центра станет вполне реальной. Такой подход выведет практику работы с дотационными регионами на высокий научный уровень.

Библиографические ссылки

1. Таштамиров МР, Байсаева МУ, Баташев РВ. Систематизация факторов и условий высокой дотационности региональных бюджетов. *Фундаментальные исследования*. 2020;11:185–192. DOI: 10.17513/fr.42896.

2. Энеева МН, Ульбашева АР, Уянаева ХБ. Факторы и причины дотационности региональных экономик СКФО. *Terra economicus*. 2010;4(3):173–176.

3. Алимурадов МК, Мидов АЗ, Одинцов СВ. Стратегический анализ бюджетной обеспеченности высокодотационных регионов. *Экономическое возрождение России*. 2021;2:114–129. DOI: 10ю37930/1990-9780-2021-2-68-113-129.

4. Кричевский МЛ, Мартынова ЮА. Использование методов машинного обучения для оценки инвестиционной деятельности различных регионов России. *Вопросы инновационной экономики*. 2019;9(4):1557–1572. DOI: 10.18334/vinec.9.4.41432.

5. Chandrinos SK, Sakkas G, Lagaros ND. AIRMS: a risk management tool using machine learning. *Expert Systems with Applications*. 2018;105:34–48. DOI: 10.1016/j.eswa.2018.03.044.

6. Stanula P, Ziegenbein A, Metternich J. Machine learning algorithms in production: a guideline for efficient data source selection. *Procedia CIRP*. 2018;78:261–266.

7. Portugal I, Alencar P, Cowan D. The use of machine learning algorithms in recommender systems: a systematic review. *Expert Systems with Applications*. 2018;97:205–227. DOI: 10.1016/j.eswa.2017.12.020.

8. Борисова ЛР, Кузнецова АВ. Использование работающего компьютерного тренажера Data Master Azforus для обучения методам машинного обучения. В: Королькова ИА, редактор. *Цифровая трансформация социальных и экономических систем. Материалы Международной научно-практической конференции; 28 января 2022 г.; Москва, Россия.* Москва: Московский университет имени С. Ю. Витте; 2022. с. 264–270.

9. Кирилюк ИЛ, Кузнецова АВ, Сенько ОВ. Исследование взаимосвязи производственных функций и социально-экономических показателей российских регионов методом оптимальных разбиений. *Информационные технологии и вычислительные системы*. 2021;1:20–31. DOI: 10.14357/20718632210103.

10. Сенько ОВ, Кузнецова АВ, Воронин ЕМ, Кравцова ОА, Борисова ЛР, Кирилюк ИЛ и др. Методы интеллектуального анализа данных в исследованиях COVID-19. *Журнал Белорусского государственного университета. Математика. Информатика*. 2022;1:83–96. DOI: 10.33581/2520-6508-2022-1-83-96.

11. Кузнецова АВ, Сенько ОВ. Возможности использования методов Data Mining при медико-лабораторных исследованиях для выявления закономерностей в массивах данных. *Врач и информационные технологии*. 2005;1:49–56.

12. Senko OV, Kuznetsova AV, Malygina NA, Kostomarova IV. Method for evaluating of discrepancy between regularities systems in different groups. *Information Technologies & Knowledge*. 2011;5(1):46–54.

Статья поступила в редколлегию 18.10.2024.
Received by editorial board 18.10.2024.