Управление в социально-экономических системах

УДК 004.852 **DOI: 10.14529/ctcr210411**

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ ПРИРОДНО-РЕСУРСНОГО ПОТЕНЦИАЛА ПЕРМСКОГО КРАЯ

А.В. Копотева[1], А.А. Максимов[2], Н.А. Сиротина[1]

[1] Пермский национальный исследовательский политехнический университет, Березниковский филиал, г. Березники, Россия,
[2] Государственная Дума Федерального Собрания Российской Федерации VIII созыва, г. Москва, Россия

В статье рассматривается проблема повышения качества моделирования и прогнозирования комплексного показателя природно-ресурсного потенциала региона за счет использования некоторых моделей машинного обучения с учителем. Актуальность решаемой задачи объясняется тем, что традиционно используемые для данных целей модели демонстрируют либо слишком низкое качество, либо сложны в настройке и оценке их параметров. **Цель исследования:** определение моделей машинного обучения, обеспечивающих оптимальные значения различных метрик качества моделирования. **Материалы и методы.** Для целей исследования рассмотрены модели множественной линейной регрессии, дерева принятия решений, случайного леса, градиентного бустинга и многослойного персептрона. В качестве метрик качества выбраны коэффициент детерминации R^2, арифметический квадратный корень из средней квадратической ошибки моделирования *RMSE*, средняя абсолютная ошибка моделирования *MAE* и относительная погрешность прогнозирования на 1 и 2 временных интервала. Исследование выполнено на примере зависимости комплексного показателя природно-ресурсного потенциала Пермского края от системы определяющих его факторов на временном интервале с 2001 по 2018 г. в среде *Jupiter Notebook* средствами библиотек *Pandas* и *Scikit-learn*. Для обеспечения сопоставимости результатов моделирования был произведен отбор факторов на основании их корреляционного анализа. Подбор оптимальных параметров моделей произведен на основании данных с 2001 по 2016 г., качество прогнозирования проверялось по данным 2017 и 2018 гг. **Результаты.** По результатам проведенного исследования оказалось, что модель классической множественной линейной регрессии демонстрирует худшие результаты по всем рассмотренным метрикам качества. Наибольшее значение коэффициента детерминации, минимальные значения корня из средней квадратичной и средней абсолютной ошибки моделирования демонстрирует модель дерева решений. При этом минимальная относительная погрешность прогнозирования на 2017 г. обеспечивается моделью градиентного бустинга, на 2018 г. – моделью многослойного персептрона. **Заключение**. Проведенное исследование позволяет утверждать, что нелинейные модели машинного обучения для задачи моделирования и прогнозирования комплексного показателя природно-ресурсного потенциала демонстрируют лучшие аппроксимационные и прогностические свойства по сравнению с множественной линейной регрессией и могут быть использованы для повышения качества управления природными ресурсами.

Ключевые слова: машинное обучение, метрики качества, регрессионный анализ, природно-ресурсный потенциал, Пермский край.

Введение

Природно-ресурсный потенциал (ПРП) региона является одной из важнейших составляющих экономического потенциала региона в целом. Качественное прогнозирование ПРП в зависимости от определяющих его факторов позволяет принимать эффективные управленческие решения при разработке планов развития территории, поэтому вопросам моделирования и прогнозирования

как показателя в целом, так и его составляющих посвящено достаточно большое число отечественных и зарубежных исследований.

Ряд работ посвящен определению и изучению комплексного показателя ПРП [1–3]. Данная группа исследований носит преимущественно качественный экономический характер, а их результаты малопригодны для количественной оценки уровня и динамики ПРП.

Для изучения отдельных составляющих ПРП широко используются различные математические методы. В силу простоты построения и удобства интерпретации часто применяются множественные линейные регрессионные модели, которые естественно использовать в качестве базы для оценки качества альтернативных моделей. В частности, в [4] сравниваются модели случайного леса и множественной линейной регрессии для оценки активности внеклеточных ферментов почв Китая.

Высокое качество прогнозирования и возможность учета скорости изменения моделируемых процессов во времени являются преимуществами использования регрессионно-дифференциальных моделей. Данный подход использован в [5] для прогнозирования перспектив горнодобывающей промышленности Пермского края. Однако оценка параметров моделей данного типа является сложным процессом с большим количеством настроек, часть из которых определяется экспериментально, поскольку не имеет теоретического или эмпирического обоснования, а практическая реализация метода требует специального программного обеспечения.

Если объект исследования может быть описан с помощью обыкновенных дифференциальных уравнений и их систем, соответствующий математический аппарат также активно используется для моделирования элементов ПРП. В частности, в [6] авторы изучают вопрос управления морскими природными ресурсами в рамках теории устойчивости систем нелинейных дифференциальных уравнений.

Широко распространено в прикладных исследованиях составляющих ПРП использование геоинформационных систем (ГИС). В частности, в работе [7] авторы применяют соответствующий аппарат для комплексной оценки природного богатства Пермского края. Исследование [8] посвящено изучению возможностей применения ГИС для управления природными ресурсами различных типов на территории Индии. Оценка запасов подземных водных ресурсов в Арабской Республике Египет с использованием геоинформационных систем и математического моделирования выполнена в [9].

При этом в силу сложности объекта исследования использование классических математических методов не всегда обеспечивает требуемое качество прогнозов.

Одним из перспективных и активно развивающихся направлений искусственного интеллекта является машинное обучение для задач регрессии и классификации. Методы машинного обучения успешно применяются для решения задач прогнозирования различных составляющих ПРП. В частности, в [10] авторы рассматривают возможности использования соответствующего аппарата в сельском хозяйстве Индии: для подбора и планирования почв, орошения и удобрения, борьбы с болезнями и вредителями. В [11] авторы применяют нейросетевые и ансамблевые алгоритмы для построения гибридной модели оценки совокупных водных запасов в бассейне реки Тебриз в северо-западном Иране. В [12] рассматривается вопрос применения гибридных алгоритмов машинного обучения для прогнозирования возникновения и распространения лесных пожаров на севере Марокко. Исследование [13] посвящено вопросу применения моделей машинного обучения для прогнозирования урожаев разных типов зерновых в разных странах. В [14] авторы проводят сравнение классических и машинных моделей добычи сырой нефти в Нигерии. Тем не менее нам не удалось обнаружить ни одного исследования, посвященного применению моделей машинного обучения для прогнозирования совокупного природно-ресурсного потенциала территории, что и определяет актуальность данной работы. Наши собственные исследования ПРП Пермского края [15] базируются на множественной линейной модели и ее модификациях, которые не обеспечивают достаточного качества прогнозирования. Рассмотрим возможность использования некоторых методов машинного обучения с учителем для повышения качества прогнозирования уровня природно-ресурсного потенциала Пермского края. Для этого необходимо:

– выбрать модели машинного обучения;

– собрать и подготовить данные;

– разбить выборку на обучающую и валидационную;
– обучить выбранные модели на обучающей выборке;
– проверить качество прогнозирования на валидационной выборке.

1. Краткая характеристика методов машинного обучения с учителем

Алгоритмы машинного обучения с учителем классифицируются в зависимости от используемого математического аппарата следующим образом[1]:
– линейные алгоритмы;
– нелинейные алгоритмы;
– ансамблевые алгоритмы;
– алгоритм искусственной нейронной сети.

Для решения задачи прогнозирования ПРП Пермского края выберем алгоритмы различных типов. Так, примером линейного алгоритма машинного обучения является обычная множественная линейная регрессия, дерево принятия решений входит в группу нелинейных алгоритмов, модели случайного леса и градиентного бустинга – примеры ансамблевых алгоритмов, а многослойный персептрон является нейронной сетью.

Коротко охарактеризуем каждый из перечисленных выше алгоритмов обучения с учителем.

1. Уравнение множественной линейной регрессии в общем случае имеет вид

$$Y_{\text{расч}}(t_k) = a + b_1 \cdot X_1(t_k) + b_2 \cdot X_2(t_k) + \ldots + b_p \cdot X_p(t_k) = a + \sum_{j=1}^{p} b_j \cdot X_j(t_k),$$

где t_k – момент времени $k = 1, 2, \ldots, K$; $Y_{\text{расч}}(t_k)$ – расчетное значение моделируемой величины в момент времени t_k; a – постоянная регрессии, определяющая уровень моделируемой величины при нулевых значениях факторов; $X_j(t_k)$ – значения факторов, определяющих значение моделируемой величины, в момент времени t_k, $j = 1, 2, \ldots, p$; b_j – коэффициенты регрессии, показывающие, насколько изменится моделируемая величина при увеличении соответствующего фактора $X_j(t_k)$ на 1.

Для определения коэффициентов модели по статистическим данным – набору векторов вида $(Y(t_k), X_1(t_k), X_2(t_k), \ldots, X_p(t_k))$, $k = 1, 2, \ldots, K$ – минимизируется сумма квадратов отклонений фактических и модельных значений вида

$$S = \sum_{k=1}^{K} \left(Y(t_k) - Y_{\text{расч}}(t_k) \right)^2 = \sum_{k=1}^{K} \left(Y(t_k) - a - \sum_{j=1}^{p} b_j \cdot X_j(t_k) \right)^2.$$

Преимуществом множественной линейной регрессии являются простота реализации и интерпретации, наличие обоснованного математического аппарата для оценки качества, недостатком – применимость модели лишь при наличии линейного тренда.

2. Дерево принятия решений (CART, Classification and Regression Trees) – двоичная рекурсивная непараметрическая процедура, позволяющая обрабатывать количественные и качественные входные и выходные величины в их исходной, необработанной форме [16]. При обучении генерируется дерево неограниченной глубины, после чего в результате различных процедур происходит отсечение части ветвей на основании анализа уровня ошибки обучения. Процедура формирования дерева инвариантна относительно порядка факторов в обучающей выборке. В результате процедуры обучения генерируется не одно, а несколько деревьев, лучшее из которых выбирается в процессе валидации на независимых данных. Разбиение в дереве решений происходит на основании правил вида «Выбираем левую ветку, если выполняется некоторое логическое условие, иначе выбираем правую ветку». Условие для разбиения подбирается исходя из минимума примеси Джини, представляющей собой вероятность неверной классификации случайно выбранного образца из некоторого их набора. Несомненными преимуществами дерева решений являются его универсальность и естественность формирования, недостатками – сложность интерпретации в случае большой глубины, а также недостаточная эффективность для решения задачи экстраполяции данных.

[1]Блиц-проверка алгоритмов машинного обучения: скорми свой набор данных библиотеке scikit-learn https://habr.com/ru/post/475552/

3. Модель случайного леса представляет собой ансамблевый алгоритм, набор решающих деревьев CART [17]. Для формирования каждого дерева в ансамбле реализуется процедура бэггинга – случайный отбор с повторениями элементов обучающей выборки в обучающую подвыборку. Каждое дерево в ансамбле строится путем случайного выбора в каждой вершине неполного набора объясняющих факторов для разбиения и генерации лучшего разбиения обучающей подвыборки на основании этих факторов. К преимуществам моделей случайного леса относят быстроту и простоту реализации, высокое качество прогнозирования и возможность обработки большого числа факторов без переобучения. К недостаткам модели относят ее большие размеры (вследствие необходимости хранить в памяти набор решающих деревьев), отсутствие четкой логики формирования результатов, а также, как и в случае отдельного решающего дерева, недостаточную эффективность для решения задачи экстраполяции данных.

4. Модель градиентного бустинга [18] также является ансамблевым алгоритмом. В отличие от модели случайного леса, в которой преимущество достигается за счет простого усреднения решений отдельных элементов ансамбля, здесь происходит последовательное добавление дополнительных элементов таким образом, чтобы получать как можно более точную оценку моделируемой величины. Иными словами, новый элемент в ансамбль включается исходя из условия максимальной корреляции с вектором антиградиента совокупной функции ошибки, связанной со всем ансамблем. При этом выбор функции ошибки производится исследователем, это может быть как классическая сумма квадратов отклонений фактических и модельных значений отклика, так и иные ее виды для непрерывных и категориальных значений отклика. Преимуществом метода градиентного бустинга является его гибкость и универсальность для решения различных прикладных задач, недостатком – большой объем используемой памяти и относительно медленная работа обученной модели.

5. Многослойный персептрон является наиболее распространенным типом нейронной сети [19]. Его отличительными особенностями являются наличие последовательных связей между слоями и отсутствие циклов и связей нейронов внутри слоев, что обеспечивает прямое распространение сигналов (от входа к выходу). Преимуществами использования нейронных сетей являются возможность решения задач при неизвестных зависимостях входных и выходных данных, устойчивость к входным возмущениям, адаптация к изменяющимся условиям среды, быстродействие за счет параллельной обработки данных и отказоустойчивость. Недостатками нейронных сетей являются необходимость экспериментального подбора структуры сети под решаемую задачу, возникновение тупиковых ситуаций при обучении, непредсказуемость результатов.

Таким образом, каждая из моделей машинного обучения имеет свои преимущества и недостатки, а подбор наиболее подходящей модели для решения конкретной задачи может быть осуществлен лишь эмпирически.

2. Исходные данные для решения задачи моделирования ПРП Пермского края

Рассмотрим возможность применения методов машинного обучения с учителем для прогнозирования природно-ресурсного потенциала Пермского края. На основании анализа региональной официальной статистики и документации был сформирован следующий перечень составляющих ПРП

– добыча нефти, включая газовый конденсат, тыс. т (Y_1);

– добыча природного и попутного газов, млн м^3 (Y_2);

– производство удобрений минеральных или химических (в пересчете на 100 % питательных веществ), тыс. т (Y_3);

– продукция сельского хозяйства, млн руб. (Y_4);

– производство деловой древесины, тыс. плотных м^3 (Y_5).

Комплексный показатель ПРП был сформирован в форме взвешенной суммы его отдельных составляющих исходя из максимальной гладкости комплексного критерия [20]. Большему значению ранга соответствует меньший вклад соответствующего частного критерия в комплексный. Полученное уравнение имеет вид $Y = Y_1/5 + Y_2/4 + Y_3/2 + Y_4/3 + Y_5/1$, т. е. наибольший вклад в ПРП Пермского края вносит производство деловой древесины, вторым по значимости является производство удобрений, третьим – производство сельскохозяйственной продукции, а наименее значимыми составляющими оказываются добыча углеводородного сырья.

Анализ статистической информации из открытых источников позволил сформировать следующий набор факторов, определяющих уровень интересующего нас показателя:

– посевные площади сельскохозяйственных культур, тыс. га (X_1);
– среднегодовая численность занятых в экономике, тыс. чел. (X_2);
– инвестиции в основной капитал предприятий, млн руб. (X_3);
– внесение удобрений минеральные удобрения, тыс. т (X_4);
– внесение удобрений органические удобрения, тыс. т (X_5);
– стоимость основных фондов на конец года, млн руб. (X_6);
– лесовосстановление, тыс. г. (X_7);
– число предприятий и организаций на конец года (X_8);
– цена нефти Brent (среднегодовое значение), USD за баррель (X_9);
– экспорт продукции ТЭК, млн USD (X_{10}).

На основании статистических данных о значения факторов (табл. 1) выполним анализ матрицы парных коэффициентов корреляции (табл. 2) и сделаем вывод о целесообразности включения объясняющих переменных в модель.

Таблица 1

Значения показателя природно-ресурсного потенциала Пермского края и определяющих его факторов в период с 2001 по 2018 г.

Table 1

Perm region natural resource potential and its factors for the period from 2001 to 2018

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
14 008,7	1237,2	1372,0	37 977,0	18,1	1,3	472 286,5	25,2	47 580,0	24,9	941,8
13 199,6	1196,9	1392,7	38 894,0	13,9	1,3	594 356,0	24,9	50 409,0	25,1	1105,3
13 564,6	1115,0	1346,0	39 679,0	12,7	1,2	683 649,0	25,3	54 612,0	28,5	1074,1
15 030,5	1061,9	1344,5	50 973,0	13,5	1,4	724 508,0	25,9	54 616,0	38,0	923,0
15 635,4	999,5	1318,9	56 800,0	13,6	1,7	961 938,0	26,4	61 002,0	55,2	353,4
16 011,6	959,5	1333,8	75 519,0	13,7	1,5	1 113 976,0	25,8	61 395,0	66,1	372,6
16 934,6	935,3	1343,4	122 480,0	15,4	1,2	1 278 827,0	25,2	58 860,0	72,7	235,7
18 362,0	914,0	1339,1	152 363,0	13,5	1,3	1 502 190,0	25,5	65 761,0	98,5	374,9
17 311,5	867,7	1316,2	134 469,0	15,3	1,4	1 605 119,0	21,2	70 784,0	62,7	144,5
18 645,9	785,5	1295,5	139 652,0	15,1	1,5	1 837 184,0	22,9	75 714,0	87,6	321,6
23 759,4	781,3	1318,9	144 781,0	16,9	1,6	2 078 245,0	28,0	77 304,0	106,4	2435,8
21 558,5	726,5	1298,7	162 241,0	17,5	1,6	2 199 176,0	26,8	75 205,0	111,7	2655,0
21 397,0	718,5	1280,1	219 494,0	14,4	1,7	2 410 614,0	30,8	77 551,0	108,7	2850,7
23 413,6	719,0	1262,0	207 597,0	15,3	1,8	2 651 647,0	27,6	76 730,0	99,5	3153,8
24 349,2	734,9	1201,0	226 214,0	13,1	2,0	2 900 859,0	32,5	83 833,0	53,6	1618,7
24 594,8	742,2	1204,4	239 390,0	13,8	2,1	3 204 554,0	29,4	81 445,0	45,2	1022,4
23 656,7	753,6	1164,5	245 140,0	17,0	2,4	3 397 061,0	38,9	76 436,0	54,8	1300,3
24 435,7	754,5	1155,6	238 008,0	16,3	2,5	3 576 306,0	40,6	70 180,0	71,6	1470,1

Таблица 2

Корреляционная матрица факторов природно-ресурсного потенциала Пермского края

Table 2

Perm Region natural resource potential factors correlation matrix

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y	1,00										
X_1	–0,91	1,00									
X_2	–0,85	0,75	1,00								
X_3	0,94	–0,91	–0,88	1,00							
X_4	0,28	–0,18	–0,15	0,20	1,00						
X_5	0,80	–0,65	–0,95	0,79	0,21	1,00					
X_6	0,96	–0,88	–0,95	0,97	0,24	0,89	1,00				
X_7	0,67	–0,48	–0,84	0,68	0,23	0,89	0,77	1,00			
X_8	0,91	–0,96	–0,74	0,88	0,10	0,63	0,86	0,41	1,00		
X_9	0,53	–0,71	–0,17	0,49	0,28	0,11	0,41	0,06	0,60	1,00	
X_{10}	0,57	–0,49	–0,30	0,46	0,31	0,33	0,46	0,36	0,47	0,50	1,00

Очевидно, что фактор X_2 связан сильной линейной зависимостью с X_3, X_5, X_6 и X_7 и в связи с этим должен быть исключен из модели. Аналогично фактор X_6 связан сильной линейной зависимостью с X_1, X_2, X_3 и X_5, его также следует исключить из рассмотрения. И, наконец, X_8 связан сильной линейной зависимостью с X_1, X_2, X_3 и X_6 и также не может входить в состав факторов. Таким образом, состав объясняющих переменных для моделирования уровня ПРП Пермского края определяется набором X_1, X_3, X_4, X_5 X_7, X_9 и X_{10}.

В качестве обучающей выборки выберем данные за период с 2001 по 2016 г., в качестве валидационной – данные за 2017 и 2018 гг. Моделирование произведем по исходным данным, поскольку практика показала, что при использовании стандартизованных данных результат оказывается значительно хуже.

3. Практическая реализация и метрики качества методов машинного обучения в задаче моделирования ПРП Пермского края

Реализацию выбранных алгоритмов машинного обучения выполним средствами Python 3.8.5 и библиотеки scikit-learn 0.23.2 в среде Jupiter Notebook 6.1.4. Для обеспечения максимального качества моделирования выполним подбор параметров алгоритмов (табл. 3), для воспроизводимости результатов зафиксируем параметр random_state=123.

Таблица 3

Оптимальные параметры моделей машинного обучения scikit-learn

Table 3

Optimal parameters of scikit-learn machine learning models

№	Модель	Параметры
1	LinearRegression	–
2	DecisionTreeRegressor	max_depth=4
3	RandomForestRegressor	n_estimators=14, max_depth=5
4	GradientBoostingRegressor	n_estimators=24
5	MLPRegressor	hidden_layer_sizes=[l=11, n=14], l – число скрытых слоев, n – число нейронов в скрытом слое

Качество моделирования будем оценивать на основании следующих метрик:

– коэффициент детерминации

$$R^2=\sum_{k=1}^{K}\left(Y_{\text{расч}}(t_k)-\overline{Y}\right)^2 \Big/ \sum_{k=1}^{K}\left(Y(t_k)-\overline{Y}\right)^2$$

характеризует долю разброса моделируемой величины, объясненной моделью; чем ближе данное значение к 1, тем лучше качество моделирования; функция для расчета метрики sklearn.metrics.r2_score;

– арифметический квадратный корень из средней квадратической ошибки моделирования

$$RMSE=\sqrt{\sum_{k=1}^{K}\left(Y_{\text{расч}}(t_k)-Y(t_k)\right)^2 \Big/ n}$$

характеризует, насколько в среднем различаются фактические и модельные значения; чем меньше значения показателя, тем выше качество моделирования; функция для расчета метрики sklearn.metrics.mean_squared_error**0,5;

– средняя абсолютная ошибка моделирования

$$MAE=\sum_{k=1}^{K}\left|Y_{\text{расч}}(t_k)-Y(t_k)\right| \Big/ n$$

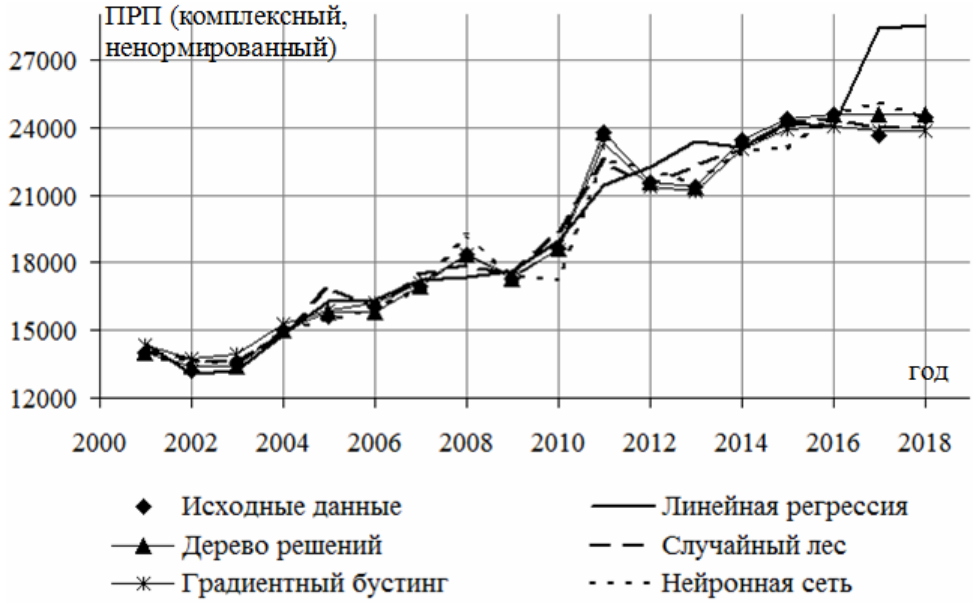
характеризует, насколько в среднем различаются фактические и модельные значения по абсолютной величине; чем меньше значения показателя, тем выше качество моделирования; функция для расчета метрики sklearn.metrics.mean_absolute_error;

– относительная погрешность прогнозирования на валидационной выборке

$$\delta(t_k)=(Y_{\text{расч}}(t_k)-Y(t_k))/Y(t_k)\cdot 100\,\%.$$

4. Анализ результатов моделирования ПРП Пермского края

На основании результатов моделирования ПРП Пермского края (см. рисунок, табл. 4) можно сделать следующие выводы.



Фактические и модельные значения природно-ресурсного потенциала Пермского края в 2001–2018 гг.
Perm region actual and model natural resource potential values in 2001–2018

Таблица 4

Метрики качества моделей природно-ресурсного потенциала Пермского края

Table 4

Perm region natural resource potential modeling quality metrics

№	Модель	R^2	*RMSE*	*MAE*	δ(2017)	δ(2018)
1	Линейная регрессия	0,828	1684,42	1040,23	19,92	16,67
2	Дерево решений	0,996	240,68	102,13	3,97	0,65
3	Случайный лес	0,980	575,95	437,19	1,48	–1,76
4	Градиентный бустинг	0,993	336,90	294,00	0,94	–2,34
5	Многослойный персептрон	0,970	701,52	483,72	6,04	–0,06

1. Классическая модель множественной линейной регрессии обеспечивает наихудшие значения всех рассмотренных метрик качества.

2. Наибольшее значение коэффициента детерминации достигается при использовании модели дерева решений, $R^2 = 0{,}996$. Чуть меньшее значение коэффициента детерминации демонстрирует модель градиентного бустинга $R^2 = 0{,}993$.

3. Наименьшее значение корня из средней квадратической ошибки моделирования также обеспечивает модель дерева решений: *RMSE* = 240,68, чуть большее значение данного показателя, как и в случае коэффициента детерминации, характерно для модели градиентного бустинга.

4. Наименьшие величины средней абсолютной ошибки моделирования также обеспечивают модели дерева решений (*MAE* = 102,13) и градиентного бустинга (*RMSE* = 294,00).

5. Минимальная относительная погрешность прогнозирования уровня ПРП в 2017 г. характерна для модели градиентного бустинга (δ(2017) = 0,94 %), чуть худший результат у модели случайного леса (δ(2017) = 1,48 %).

6. Минимальная относительная погрешность прогнозирования уровня ПРП в 2018 г. (по абсолютной величине) характерна для модели многослойного персептрона (δ(2018) = –0,06 %), чуть худший результат у модели дерева решений (δ(2018) = 0,65 %).

Заключение

Проведенный в данном исследовании анализ возможности применения моделей машинного обучения с учителем, отличных от множественной линейной регрессии (дерева решений CART, случайного леса, градиентного бустинга и многослойного персептрона), позволяет не только обеспечить лучшие значения метрик качества, но и значительно повысить точность прогнозирования комплексного показателя ПРП Пермского края.

Литература

1. Samus, T. Assessing the natural resource use and the resource efficiency potential of the Desertec concept / T. Samus, B. Lang // Solar Energy. – 2013. – Vol. 87. – P. 176–183.

2. Березовский, П.В. Экономическая оценка вторичных минеральных ресурсов: моногр. / П.В. Березовский. – СПбГГИ им. Г.В. Плеханова, 2009. – 161 с.

3. Соколова, Н.В. Природно-ресурсный потенциал территории: содержание понятия, методы оценки / Н.В. Соколова // Вестник Ленинградского университета. – 1988. – № 3. – С. 125–130.

4. Xie, X. Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land / X. Xie, T. Wu, M. Zhu et al. // Ecological Indicators. – 2021. – Vol. 120, 106925.

5. Ракаева, Т.Г. Регрессионно-дифференциальная модель динамики горной промышленности Пермского края / Т.Г. Ракаева, В.Ф. Беккер // Системный анализ в науке и образовании. – 2019. – № 2. – С. 45–51.

6. Biswas, M.H. Mathematical Modeling Applied to Sustainable Management of Marine Resources / M.H. Biswas, M.R. Hossain, M.K. Mondal // Procedia Engineering. – 2017. – Vol. 194. – P. 337–344.

7. Красильников, П.А. Геоинформационное обеспечение экономической оценки природно-ресурсного потенциала территорий Пермского края / П.А. Красильников // Экономика региона. – 2009. – № 1. – С. 143–151.

8. Kumar, N. Applications of Remote Sensing and GIS in Natural Resource Management / N. Kumar, S.S. Yamac, A. Velmurugan // Journal of the Andaman Science Association. – 2015. – Vol. 20 (1). – P. 1–6.

9. Multi-criteria decision support for geothermal resources exploration based on remote sensing, GIS and geophysical techniques along the Gulf of Suez coastal area, Egypt / S. Abuzied, M. Kaiser, E. Shendi, M. Abdel-Fattah // Geothermics. – 2020. – Vol. 88, 101893.

10. Akhter, R. Precision agriculture using IoT data analytics and machine learning / R. Akhter, S.A. Sofi // Journal of King Saud University – Computer and Information Sciences. – 2021, 101016.

11. Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques / A. Arabameri, S.C. Pal, F. Rezaie et al. // Journal of Hydrology: Regional Studies. – 2021. – Vol. 36, 100848.

12. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area / M. Mohajane, R. Costache, F. Karimi et al. // Ecological Indicators. – 2021. – Vol. 129, 107869.

13. Machine learning for large-scale crop yield forecasting / D. Paudel, H. Boogaard, A. de Wit et al. // Agricultural Systems. – 2021. – Vol. 187, 103016.

14. Classical and machine learning modeling of crude oil production in Nigeria: Identification of an eminent model for application / C.P. Obite, A. Chukwu., D.C. Bartholomew et al. // Energy Reports. – 2021. – Vol. 7. – P. 3497–3505.

15. Сиротина, Н.А. Применение конечно-разностных моделей для краткосрочного прогнозирования природно-ресурсного потенциала Пермского края / Н.А. Сиротина, А.В. Копотева, А.В. Затонский // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2021. – Т. 21, № 2. – С. 154–166. DOI: 10.14529/ctcr210215

16. Wu, X. Top 10 algorithms in data mining / X. Wu, V. Kumar, Quinlan R. et al. // Knowledge and Information Systems. – 2008. – Vol. 14. – P. 1–37.

17. Biau, G. Analysis of a Random Forests Model / G. Biau // Journal of Machine Learning Research. – 2012. – Vol. 13. – P. 1063–1095.

18. Natekin, A. Gradient Boosting Machines, A Tutorial / A. Natekin, A. Knoll // Frontiers in neurorobotics. – 2013. – Vol. 7. – P. 21.

19. Multilayer perceptron and neural networks / P. Marius, V. Balas, L. Perescu-Popescu, N. Mastorakis // WSEAS Transactions on Circuits and Systems. – 2009. – Vol. 8. – P. 579–588.

20. Сиротина, Н.А. Оценка вклада горнодобывающей отрасли в природно-ресурсный потенциал региона / Н.А. Сиротина, А.В. Копотева, А.В. Затонский // Горный информационно-аналитический бюллетень (научно-технический журнал). – 2020. – № 8. – С. 163–178. DOI: 10.25018/0236-1493-2020-8-0-163-178

Копотева Анна Владимировна, канд. техн. наук, доцент кафедры общенаучных дисциплин, Пермский национальный исследовательский политехнический университет, Березниковский филиал, г. Березники; kopoteva_av@mail.ru.

Максимов Александр Александрович, д-р техн. наук, депутат, Государственная Дума Федерального Собрания Российской Федерации VIII созыва, г. Москва.

Сиротина Наталья Александровна, старший преподаватель кафедры общенаучных дисциплин, Пермский национальный исследовательский политехнический университет, Березниковский филиал, г. Березники; nsirotina117@mail.ru.

Поступила в редакцию 24 сентября 2021 г.

DOI: 10.14529/ctcr210411

PERM REGION NATURAL RESOURCE POTENTIAL FORECASTING USING MACHINE LEARNING MODELS

A.V. Kopoteva*[1]*, kopoteva_av@mail.ru,
A.A. Maksimov*[2]*,
N.A. Sirotina*[1]*, nsirotina117@mail.ru
[1] Perm National Research Polytechnic University, Berezniki Branch, Berezniki, Russian Federation,
[2] State Duma of the Federal Assembly of the Russian Federation of the VIII convocation, Moscow, Russian Federation

In the article we consider a complex indicator of region natural resource potential modeling and forecasting quality improvement using different machine learning models. Problem under consideration importance is determined by the fact that the models traditionally used for these purposes demonstrate either low quality, or high configuration and parameters evaluation difficulty. **The aim** of the study is determination of machine learning models that provide the optimal values of various modeling quality metrics. **Materials and methods**. For this study purposes we considered the multiple linear regression, decision tree, random forest, gradient boosting and multilayer perceptron models. We used the determination coefficient R^2, the root mean square error of modeling RMSE, the average absolute error of modeling MAE, and the relative error of prediction for 1 and 2 time intervals as quality metrics. This study is based on data of the complex indicator of the Perm Region natural resource potential and the system of its determining factors in the time interval from 2001 to 2018. We evaluate models and calculate quality metrics using Pandas and Scikit-learn Python libraries in Jupiter Notebook environment. **Results.** According to our research the classical multiple linear regression model demonstrates the worst results for all quality metrics under consideration. The decision tree model demonstrates determination coefficient maximum value and minimum root mean square error and mean absolute error. Minimum relative forecasting error for 2017 is provided by the gradient boosting model, for 2018 – by the multilayer perceptron model. **Conclusion**. Our study allows us to affirm that nonlinear machine learning models for the task of region natural resource

potential modeling and forecasting demonstrate better approximating and predictive properties compared to multiple linear regression and thus can be used to improve the quality of natural resource management.

Keywords: machine learning, quality metrics, regression analysis, natural resource potential, Perm region.

References

1. Samus T., Lang B. Assessing the natural resource use and the resource efficiency potential of the Desertec concept. *Solar Energy*, 2013, vol. 87, pp. 176–183.

2. Berezovskiy P.V. *Ekonomicheskaya otsenka vtorichnykh mineral'nykh resursov: monogr.* [Economic assessment of secondary mineral resources: monograph]. SPbGGI named after G. V. Plekhanov, 2009. 161 p.

3. Sokolova N.V. [Natural resource potential of the territory: the content of the concept, methods of assessment]. *Bulletin of the Leningrad University*, 1988, no. 3. pp. 125–130. (in Russ.)

4. Xie X., Wu T., Zhu M., Jiang G., Xu W. Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *EcologicalIndicators*, 2021, Vol. 120, 106925.

5. Rakaeva T.G., Bekker V.F. [Regression-differential model of the dynamics of the mining industry of the Perm Region]. *System analysis in science and education*, 2019, no. 2. pp. 45–51. (in Russ.)

6. Biswas M.H., Hossain M.R., Mondal M.K. Mathematical Modeling Applied to Sustainable Management of Marine Resources. *Procedia Engineering*, 2017, vol. 194, pp. 337–344.

7. Krasil'nikov P.A. [Geoinformation support for the economic assessment of the natural resource potential of the Perm Krai territories]. *Economy of the region*, 2009, no. 1, pp. 143–151. (in Russ.)

8. Kumar N. Yamas S.S., Velmurugan A. Applications of Remote Sensing and GIS in Natural Resource Management. *Journal of the Andaman Science Association*, 2015, vol. 20 (1), pp. 1–6.

9. Abuzied S., Kaiser M., Shendi E., Abdel-Fattah M. Multi-criteria decision support for geothermal resources exploration based on remote sensing, GIS and geophysical techniques along the Gulf of Suez coastal area, Egypt. *Geothermics*, 2020, vol. 88, 101893.

10. Akhter R., Sofi S.A. Precision agriculture using IoT data analytics and machine learning. *Journal of King Saud University – Computer and Information Sciences*, 2021, 101016.

11. Arabameri A., Pal S.C., Rezaie F., Nalivan O.A., Chowdhuri I., Saha A., Lee S., Moayedi H. Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques. *Journal of Hydrology: Regional Studies*, 2021, vol. 36, 100848.

12. Mohajane M., Costache R., Karimi F., Pham Q.B., Essahlaoui A., Nguyen H., Laneve G., Oudija F. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. *Ecological Indicators*, 2021, vol. 129, 107869.

13. Paudel D., Boogaard H., de Wit A., Janssen S., Osinga S., Pylianidis C., Athanasiadis I.N. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 2021, vol. 187, 103016.

14. Obite C.P., Chukwu A., Bartholomew D.C., Nwosu U.I., Esiaba G.E. Classical and machine learning modeling of crude oil production in Nigeria: Identification of an eminent model for application. *Energy Reports*, 2021. vol. 7, p. 3497–3505.

15. Sirotina N.A., Kopoteva A.V., Zatonskiy A.V. Finite-Difference Models Application for Short-Term Forecasting of the Natural Resource Potential of the Perm Region. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2021, vol. 21, no. 2, pp. 154–166. (in Russ.) DOI: 10.14529/ctcr210215

16. Wu X., Kumar V., Quinlan R., Ghosh J., Yang Q. Motoda H., Mclachlan G., Ng S.K.A., Liu B., Yu P., Zhou Z.-H., Steinbach M., Hand D., Steinberg D. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 2008, vol. 14, pp. 1–37.

17. Biau G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 2012, vol. 13, pp. 1063–1095.

18. Natekin A., Knoll A. Gradient Boosting Machines, A Tutorial. *Frontiers in neurorobotics*, 2013, vol. 7, p. 21.

19. Marius P., Balas V., Perescu-Popescu L., Mastorakis N. Multilayer perceptron and neural networks. *WSEAS Transactionson Circuits and Systems*, 2009, vol. 8, pp. 579–588.

20. Sirotina N.A., Kopoteva A.V., Zatonskiy A.V. [Assessment of the contribution of the mining industry to the natural resource potential of the region]. *Gornyy informatsionno-analiticheskiy byulleten' (nauchno-tekhnicheskiy zhurnal)* [Mining Information and Analytical Bulletin (scientific and technical journal)], 2020, no. 8, pp. 163–178. (in Russ.) DOI: 10.25018/0236-1493-2020-8-0-163-178

Received 24 September 2021

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Копотева, А.В. Модели машинного обучения в задаче прогнозирования природно-ресурсного потенциала Пермского края / А.В. Копотева, А.А. Максимов, Н.А. Сиротина // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2021. – Т. 21, № 4. – С. 126–136. DOI: 10.14529/ctcr210411

FOR CITATION

Kopoteva A.V., Maksimov A.A., Sirotina N.A. Perm Region Natural Resource Potential Forecasting Using Machine Learning Models. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2021, vol. 21, no. 4, pp. 126–136. (in Russ.) DOI: 10.14529/ctcr210411