Вестник Самарского университета. Экономика и управление. 2024. Т. 15. № 4. С. 187–201
doi: 10.18287/2542-0461-2024-15-4-187-201 ISSN: 2782-3008(online), ISSN 2542-0461(print)

УДК 332.142.2

Построение и применение инновационного рейтинга регионов с использованием технологии случайного леса

С. Н. Яшин, Н. И. Яшина, Е. В. Кошелев

Национальный исследовательский Нижегородский государственный университет им. Н. И. Лобачевского, Россия, 603022, г. Нижний Новгород, пр. Гагарина, 23.

Аннотация

Публикуемая статья посвящена актуальной проблеме разработки моделей построения и применения инновационного рейтинга регионов страны, имеющих отрасль радиоэлектронной промышленности (РЭП). Для преодоления ряда недостатков известных классических статистических подходов для повышения точности прогнозирования развития отрасли РЭП и адекватности присваиваемых ей рейтингов. Разработан и применен вариант технологии машинного обучения «случайный лес», повышающий точность прогнозирования развития отрасли РЭП. Проведена верификация полученных рейтингов на данных нового периода наблюдения с целью определения по регионам-лидерам сегментов входных переменных модели.

Ключевые слова: радиоэлектронная промышленность; инновационный рейтинг; задача классификации; случайный лес.

Получение: 12 октября 2024 г. / Исправление: 11 ноября 2024 г. /
Принятие: 11 декабря 2024 г. / Публикация онлайн: 28 января 2025 г.

Региональная и отраслевая экономика (научная статья)

Образец для цитирования:
Яшин С. Н., Яшина Н. И., Кошелев Е. В. Построение и применение инновационного рейтинга регионов с использованием технологии случайного леса // *Вестник Самарского университета. Экономика и управление*, 2024. Т. 15, № 4. С. 187–201. doi: http://doi.org/10.18287/2542-0461-2024-15-4-187-201.

Сведения об авторах:
Сергей Николаевич Яшин http://orcid.org/0000-0002-7182-2808
д.э.н, профессор; заведующий кафедрой менеджмента и государственного управления; e-mail: jashinsn@yandex.ru
Надежда Игоревна Яшина http://orcid.org/0000-0002-0630-7949
д.э.н, профессор; заведующая кафедрой финансов и кредита;
e-mail: yashina@iee.unn.ru
Егор Викторович Кошелев http://orcid.org/0000-0001-5290-7913
к.э.н., доцент; доцент кафедры менеджмента и государственного управления; e-mail: ekoshelev@yandex.ru

Введение

В современных условиях развития экономики существует острая необходимость планирования инновационного развития приоритетных отраслей промышленности. Отрасль радиоэлектронной промышленности (РЭП) является одной из ключевых в плане решения задачи импортозамещения.

Государственная поддержка регионов, имеющих необходимый инновационный потенциал для выполнения данной задачи, требует предварительной оценки инновационных перспектив регионов страны, имеющих отрасль РЭП. Для этого необходимо создание инновационного рейтинга регионов. Важность развития отрасли РЭП для развития общества сложно переоценить.

В настоящее время особая роль в применении продукции радиоэлектронной промышленности (РЭП) отводится оборонно-промышленному комплексу, как инструменту обеспечения вооруженных сил современным и качественным вооружением, военной и специальной техникой. Вместе с тем очевидным трендом является все большая ориентированность радиоэлектронной промышленности на гражданскую сферу [1].

Промышленная электроника является важным аспектом современного производства, обеспечивая автоматизацию, управление, мониторинг и связь в промышленных процессах. Последние тенденции в области промышленной электроники, такие как IIoT и Industry 4.0, еще больше повышают эффективность и снижают издержки, делая обрабатывающую промышленность более конкурентоспособной и устойчивой [2].

Целью исследования является построение и применение инновационного рейтинга регионов страны в отрасли РЭП.

В настоящей работе проводится построение и применения инновационного рейтинга регионов страны, имеющих отрасль РЭП. Классические статистические подходы имеют ряд недостатков, сказывающихся на точности прогнозов развития, а, следовательно, и адекватности присваиваемых рейтингов. По нашему мнению, технологии машинного обучения позволяют во многом решить задачу повышения точности прогнозов будущего развития регионов. Одной из таких технологий является «случайный лес» (Random Forest, RF).

Случайный лес обладает рядом преимуществ по сравнению с другими технологиями машинного обучения:

1. Для входных данных не требуется их стандартизация.
2. Входные данные не нужно самостоятельно разбивать на обучающую и тестовую выборки.
3. Случайный лес не склонен к переобучению, за исключением случаев сильно зашумленных данных.

При этом данная технология имеет и недостатки:

1. Входные параметры не должны быть коррелированными между собой.
2. В случае сильно зашумленных данных приходится выполнять подрезку решающих деревьев, чтобы избежать переобучения алгоритма.

Имеются также другие преимущества и недостатки случайного леса. Мы лишь перечислили основные из них.

Обсудим тогда последние достижения технологии случайного леса, а также проиллюстрируем широкие возможности применения данного алгоритма в различных областях научного знания.

Ансамблевые методы, такие как случайный лес, хорошо работают с многомерными наборами данных. Однако, когда количество признаков чрезвычайно велико по сравне-

нию с количеством выборок, а процент действительно информативных признаков очень мал, производительность традиционного случайного леса значительно снижается. С этой целью Д. Гош и Дж. Кабрера разработали новый подход, который повышает производительность традиционного случайного леса за счет уменьшения вклада деревьев, узлы которых заполнены менее информативными признаками [3].

С. Тиянга, Р. Дж. Талагала и Г. А. Хиндмен используют случайный лес для определения лучшего метода прогнозирования с использованием только функций временных рядов [4].

Система оценивается с использованием временных рядов соревнований M1 и M3 и показывает, что она дает точные прогнозы, сравнимые с некоторыми эталонными показателями и другими широко используемыми автоматизированными подходами прогнозирования временных рядов. Ключевым преимуществом предлагаемой авторами структуры является то, что трудоемкий процесс создания классификатора выполняется до решения текущей задачи прогнозирования.

Деревья решений, используемые для построения случайного леса, могут иметь низкую точность классификации или высокую корреляцию, что влияет на комплексную производительность случайного леса. Стремясь решить эти проблемы, в Ж. Сан и др. предложили улучшенный случайный лес, основанный на точности классификации и измерении корреляции деревьев решений [5].

Уделяя особое внимание задачам классификации, М. Сиппер и Дж. Г. Мур провели обширные эксперименты с сохранением случайных лесов, включая 5 методов выращивания (включая представленный авторами новый – лексигарден), 6 источников наборов данных и 31 набор данных [6].

Ученые показали, что значительного улучшения можно достичь, используя модели, которыми в любом случае уже располагаем, и предусматриваем возможность создания хранилищ моделей (а не просто наборов данных, решений или кода), которые можно было бы сделать доступными для всех, тем самым способствуя развитию данных и вычислительной науки.

Т. Суреш и др. [7] предложили гибридную модель для прогнозирования заболеваний сердца с использованием случайного леса и машины опорных векторов [7].

При использовании случайного леса выполняется итеративное исключение признаков для выбора признаков заболеваний сердца, что улучшает прогностический результат машины опорных векторов для прогнозирования заболеваний сердца.

Эксперимент проводился с предлагаемой моделью с использованием тестового набора, и результаты эксперимента, очевидно, доказывают, что производительность предлагаемой гибридной модели лучше по сравнению с отдельной машиной случайного леса и опорных векторов. В целом авторы разработали более точную и вычислительно эффективную модель прогнозирования заболеваний сердца с точностью 98.3%.

Н. М. Абдулкарим и А. М. Абдулазиз процесс создания случайных лесов и статус исследования случайных лесов представили с точки зрения повышения потенциала и показателей эффективности [8].

Б. Лоэф, А. Вонг и Н. А. Г. Янссен использовали случайный лес (RF) для выявления наиболее сильных предикторов из-за его благоприятных показателей прогнозирования в предыдущих исследованиях [9].

Связь между предикторами и исходом визуализировалась с помощью графиков частичной зависимости и накопленных локальных эффектов. Для облегчения интерпретации риски были суммированы, выражая их как средний риск и среднюю тенденцию с

течением времени. Способность модели RF отличать плохие от хороших самооценок здоровья была приемлемой.

С. Вонгвибулсин, К. К. Ву и С. Л. Зегер исследовали новый метод статистики и машинного обучения RF-SLAM, который улучшает прогнозирование рисков за счет включения изменяющейся во времени информации и учета большого количества предикторов, их взаимодействий и пропущенных значений [10].

RF-SLAM предназначен для легкого расширения до одновременного прогнозирования множества конкурирующих событий и/или повторных измерений дискретных или непрерывных переменных с течением времени.

Опираясь на структуру эксперимента и проверки, предложенную в эксперименте COST action VALUE – крупнейшем и наиболее исчерпывающем исследовании методов статистического даунскейлинга на сегодняшний день – М. Н. Легаса и др. ввели и тщательно проанализировали апостериорные случайные леса (AP–RF), в которых используется вся информация, содержащаяся в листьях, позволяет достоверно прогнозировать форму и масштабные параметры гамма-распределения вероятностей осадков в дождливые дни [11].

Т. Ука, Г. Йохно и К. Накамото использовали два аналитических метода для сравнения прогностической способности: RF как новую модель и многомерную логистическую регрессию (MLR) как традиционную модель [12].

Авторы также создали модели, исключив значения изменений, чтобы определить, оказало ли это положительное влияние на прогнозы. Кроме того, в RF-анализе рассчитывалась важность переменных, а в MLR-анализе рассчитывались стандартные коэффициенты регрессии для идентификации предикторов. Модель RF показала более высокую прогностическую способность изменения HbA1c, чем MLR во всех моделях. RF-модель, включающая значения изменений, показала наибольшую предсказательную силу.

В работе Й. Гуо и др. используются методы NLP и разработка функций, а также случайный лес для идентификации слухов для достижения точной идентификации и прогнозирования слухов [13].

Исследования показали, что слухи на платформе *Weibo* можно точно идентифицировать. Путем точного выявления слухов это исследование направлено на снижение вредного воздействия слухов и предоставление обществу информации.

Дж. Клусовски возвратился к исторически важной модели случайного леса, называемой центрированными случайными лесами, первоначально предложенной Брейманом в 2004 году и позже изученной Жераром Био в 2012 году, где объект выбирается случайным образом, а расщепление происходит в средней точке узла вдоль выбранного объекта [14].

Более того, путем детального анализа ошибок аппроксимации и оценки для линейных моделей автор показал, что полученную новую скорость в целом невозможно улучшить. Наконец, Дж. Клусовски улучшил текущие границы ошибок прогнозирования для другой модели случайного леса, называемой медианными случайными лесами, в которой каждое дерево строится на основе подвыборки данных, а разбиение выполняется по эмпирической медиане вдоль выбранного признака.

М. А. Хасан Далфи, С. Чабоуни и А. Фахфах предложили инновационную методологию обнаружения рака молочной железы, основанную на машинном обучении, которая использует алгоритм случайного леса с помощью выбора функций [15].

Рамки исследования включают в себя передовые методы выбора признаков, такие как коэффициент инфляции дисперсии (VIF), выбор признаков на основе модели, рекурсивное исключение признаков и одномерный выбор признаков, для извлечения весьма релевант-

ных признаков и выявления скрытых закономерностей, связанных с опухолями.

1. Модель

Основная идея представленной далее модели заключается в том, что инновационный рейтинг регионов страны строится по трем целевым функциям:

1) объем инновационных товаров (всего) (млн руб.),
2) разработанные передовые производственные технологии (всего) (ед.),
3) сальдированный финансовый результат (информатизация и связь) (млн руб.).

Таким образом, хотя рейтинг строится для отрасли РЭП, он также характеризует, как данная отрасль влияет на инновационное развитие и других отраслей промышленности – по целевым функциям 1 и 2. В этом заключается его преимущество.

Сам инновационный рейтинг подразумевает деление регионов на три класса: A – регионы-лидеры, B – регионы со средним уровнем инновационного развития, C – депрессивные регионы.

Построение и применение обозначенного инновационного рейтинга регионов подразумевает применение машинного обучения, а именно, решение задачи классификации с использованием технологии случайного леса, а далее верификация полученных рейтингов на данных нового периода наблюдения с целью определения по регионам-лидерам сегментов входных переменных модели. Эти сегменты являются плановыми показателями для того, чтобы в дальнейшем определить, будет ли регион иметь инновационный рейтинг А. Это поможет государственным структурам определять, какие регионы следует поддерживать в их инновационном развитии.

На рис. 1 подробно представлены этапы построения и применения инновационного рейтинга регионов с использованием технологии случайного леса.

Этап 1 – сбор, корректировка на инфляцию и стандартизация необходимых для анализа данных. На данном этапе с сайта Федеральной службы государственной статисти-

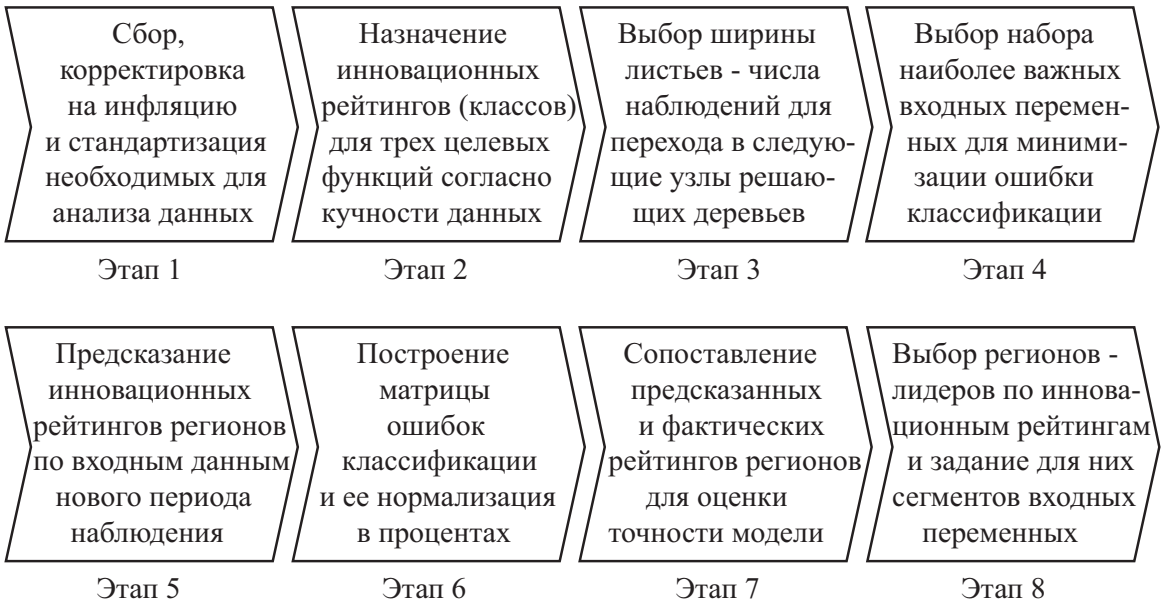


Рис. 1: Этапы построения и применения инновационного рейтинга регионов с использованием технологии случайного леса.

Fig. 1: Stages of building and applying an innovative rating of regions using random forest technology.

ки (`www.gks.ru`) собираем необходимые нам данные за период с 2010 по 2022 гг. о следующих входных переменных и целевых функциях по 83 регионам России:

1) вход 1 – стоимость основных фондов (ОФ) (информатизация и связь) (млн руб.) (x_1);
2) вход 2 – ввод в действие основных фондов (ОФ) (информатизация и связь) (млн руб.) (x_2);
3) вход 3 – степень износа основных фондов (ОФ) (информатизация и связь) (%) (x_3);
4) вход 4 – оборот организаций (информатизация и связь) (млрд руб.) (x_4);
5) вход 5 – затраты на внедрение и использование цифровых технологий (всего) (млн руб.) (x_5);
6) вход 6 – внутренние текущие затраты на НИР (фундаментальные исследования) (млн руб.) (x_6);
7) вход 7 – внутренние текущие затраты на НИР (прикладные исследования) (млн руб.) (x_7);
8) вход 8 – внутренние текущие затраты на НИР (разработки) (млн руб.) (x_8);
9) вход 9 – затраты на инновационную деятельность (всего) (млн руб.) (x_9);
10) вход 10 – используемые передовые производственные технологии (всего) (ед.) (x_{10});
11) цель 1 – объем инновационных товаров (всего) (млн руб.) (y_1);
12) цель 2 – разработанные передовые производственные технологии (всего) (ед.) (y_2);
13) цель 3 – сальдированный финансовый результат (информатизация и связь) (млн руб.) (y_3).

В результате получаем матрицу данных размерности $1\,079 \times 13$.

Для того, чтобы данные в рублях были сравнимы между собой в разные годы, их необходимо скорректировать на инфляцию. Для этого используем данные о годовой инфляции с сайта Федеральной службы государственной статистики (`www.gks.ru`). Таким образом, получаем данные в ценах последнего 2022 г.

Для реализации алгоритма случайного леса нет необходимости стандартизации полученных данных. Однако мы это делаем прежде всего для того, чтобы в дальнейшем было проще назначить инновационные рейтинги (классы) для трех целевых функций.

Этап 2 – назначение инновационных рейтингов (классов) для трех целевых функций согласно кучности данных. В программе *Statistica* строим график распределения стандартизованных данных для каждой целевой функции. Затем по этому графику смотрим, насколько кучно данные распределены в той или иной части графика. Это позволяет определить интервалы, соответствующие каждому из трех рейтингов – A, B, C.

Этап 3 – выбор ширины листьев – числа наблюдений для перехода в следующие узлы решающих деревьев. Обучаем модель на данных с 2010 по 2021 гг. Данные 2022 г. оставляем для верификации обученной модели.

В программе *Matlab* можно реализовать алгоритм случайного леса. Для этого на данном этапе выращиваем разные варианты случайного леса в зависимости от ширины листьев. При этом мы хотим минимизировать ошибку классификации по технологии “out-of-bag” (OOB). Ошибка OOB – это средняя ошибка прогнозирования для каждой обучающей выборки x_i, использующая только деревья, которые не имели x_i в своей выборке начальной загрузки.

В итоге выбираем тот вариант ширины листьев, который позволяет получить наименьшую ошибку OOB при соответствующем количестве выращенных деревьев. Здесь мы фиксируем параметр ширины листьев и количества выращенных деревьев, чтобы использовать их на следующих этапах моделирования.

Этап 4 – выбор набора наиболее важных входных переменных для минимизации ошиб-

ки классификации. Данная процедура является альтернативой анализу корреляционной матрицы. То есть она позволяет выбрать из всего набора входных переменных только те, которые оказывают наибольшее влияние на значение целевой функции (или правильность классификации). При этом выбираются те переменные, которые позволяют уменьшить ошибку OOB или же получить ее примерно такой же, как в случае использования всего набора переменных.

Этап 5 – предсказание инновационных рейтингов регионов по входным данным нового периода наблюдения. Здесь мы начинаем верификацию полученной модели. Для этого сначала прогнозируем, какой получится рейтинг для каждого из 83 регионов страны в последнем 2022 г. в случае каждой целевой функции. Результат можно получить в виде вероятностей отнесения того или иного региона к соответствующему рейтингу. В итоге выбирается рейтинг с наибольшим значением его вероятности.

Этап 6 – построение матрицы ошибок классификации и ее нормализация в процентах. Данная матрица позволяет сравнить предсказанные значения рейтингов с фактическими. На главной диагонали матрицы отмечаются совпадения предсказаний с фактическими значениями. Чем их больше по сравнению с соседними случаями несовпадений, тем качественнее обученная модель случайного леса.

Эту матрицу также можно нормализовать, чтобы увидеть совпадения и несовпадения предсказаний и фактических значений в процентах. Зачастую это бывает удобнее.

Этап 7 – сопоставление предсказанных и фактических рейтингов регионов для оценки точности модели. Здесь подсчитывается число совпадений предсказанных и фактических рейтингов по всем 83 регионам страны. Затем число совпадений определяем в процентах. Этот результат иллюстрирует точность обученной модели.

Этап 8 – выбор регионов – лидеров по инновационным рейтингам и задание для них сегментов входных переменных. Здесь в 2022 г. сначала выбираем те регионы, у которых наблюдался фактический рейтинг A хотя бы для одной целевой функции. Затем среди полученного набора регионов для каждой входной переменной находим худшее и лучшее значение. Именно эти границы и определяют сегменты входных переменных. Они являются плановыми для того, чтобы в дальнейшем определить, будет ли регион иметь инновационный рейтинг A. Таким образом, на этом последнем этапе мы переходим к непосредственному применению полученных нами инновационных рейтингов регионов.

2. Результаты

Проиллюстрируем построение и применение инновационного рейтинга регионов, используя для этого данные Федеральной службы государственной статистики (`www.gks.ru`).

Этап 1 – сбор, корректировка на инфляцию и стандартизация необходимых для анализа данных. Проводя указанные для данного этапа процедуры, получаем матрицу данных, представленную в таблице 3.

Этап 2 – назначение инновационных рейтингов (классов) для трех целевых функций согласно кучности данных. Дальнейший анализ опишем на примере цели 1 – объем инновационных товаров (всего) (млн руб.).

График кучности данных (рис. 2), полученный в программе *Statistica*, позволяет определить интервалы для каждого рейтинга:

1) рейтинг A – интервал стандартизованных значений $(1; +\infty)$;

2) рейтинг B – полуинтервал стандартизованных значений $(-0.4; 1]$;

3) рейтинг C – интервал стандартизованных значений $(-\infty; -0.4]$.

Этап 3 – выбор ширины листьев – числа наблюдений для перехода в следующие узлы решающих деревьев. Обучая модель случайного леса для цели 1 в программе *Matlab*, получаем, что наименьшая ошибка OOB получается в случае ширины листьев 10 и количестве выращенных деревьев 28 (рис. 3, *a*). Фиксируем полученные параметры, чтобы использовать их на следующих этапах моделирования.

Этап 4 – выбор набора наиболее важных входных переменных для минимизации ошибки классификации. На рис. 3, *b* видно, что наиболее важные входные переменные, которые оказывают наибольшее влияние на правильность классификации, это переменные 8–10. Однако, постепенно добавляя все остальные переменные в модель, мы можем получить наименьшее значение ошибки OOB (рис. 4, *a*). То есть используем все 10 входных переменных.

Таблица 1: Данные для построения инновационного рейтинга регионов.

Table 1: Data for constructing an innovation rating of regions.

Регион	**Вход 1**	**Вход 2**	**Вход 3**	**Вход 4**	**Вход 5**	**Вход 6**	**Вход 7**	**Вход 8**
2010								
1.Белгородская область	−0.2972	−0.2499	−1.138	−0.1493	−0.1325	−0.2175	−0.1844	−0.2689
2.Брянская область	−0.2377	−0.2045	−0.0984	−0.1013	−0.1535	−0.2479	−0.2306	−0.2875
...	...	...	...	...	...	...	...	...
83.Чукотский округ	−0.3505	−0.2234	−2.0437	−0.1601	−0.1667	−0.2585	−0.2374	−0.2932
...								

Регион	**Вход 9**	**Вход 10**	**Цель 1**	**Рейтинг**	**Цель 2**	**Рейтинг**	**Цель 3**	**Рейтинг**
2010								
1.Белгородская область	−0.2917	−0.4571	−0.3253	'B'	−0.1943	'B'	−0.0717	'B'
2.Брянская область	−0.3722	−0.5152	−0.4006	'C'	−0.31	'C'	−0.1322	'C'
...	...	...	...	...	...	...	...	...
83.Чукотский округ ский округ	−0.4067	−0.8214	−0.4651	'C'	−0.4257	'C'	−0.1283	'B'
...								

Этап 5 – предсказание инновационных рейтингов регионов по входным данным нового периода наблюдения. В таблице 2 показаны вероятности рейтингов в 2022 г. согласно предсказаниям обученной модели. Рейтинги, имеющие наибольшие вероятности, отражены в колонке "Предсказанный рейтинг".

Этап 6 – построение матрицы ошибок классификации и ее нормализация в процентах. Нормализованная матрица ошибок классификации показана на рис. 4, *b*. С наибольшей точностью 93.5% обученная модель предсказывает на новых данных 2022 г. рейтинг B. С наименьшей точностью 71.4% она предсказывает рейтинг A.

Этап 7 – сопоставление предсказанных и фактических рейтингов регионов для оценки точности модели. На основе данных таблицы 2 получается, что модель на новых данных 2022 г. не ошибается в предсказании рейтингов регионов в 72 случаях из 83. То есть точность предсказаний составляет 86.75%. Аналогично для целей 2 и 3 точность предсказаний рейтингов в 2022 г. составляет в обоих случаях 69.88%.

Этап 8 – выбор регионов – лидеров по инновационным рейтингам и задание для них сегментов входных переменных. Предварительно сначала стандартизованные данные переводим обратно в реальные (с учетом инфляции).

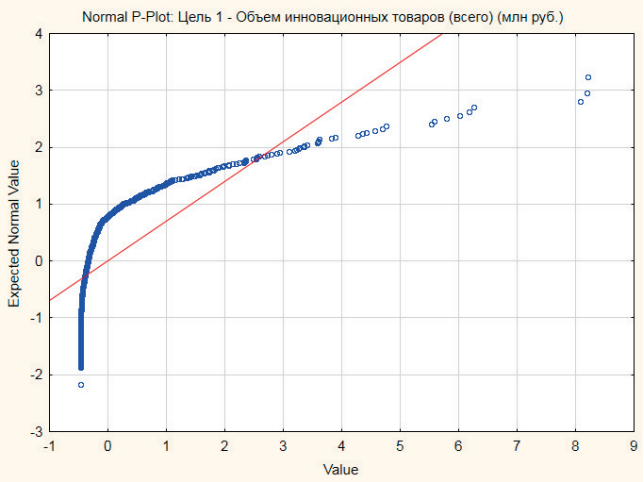
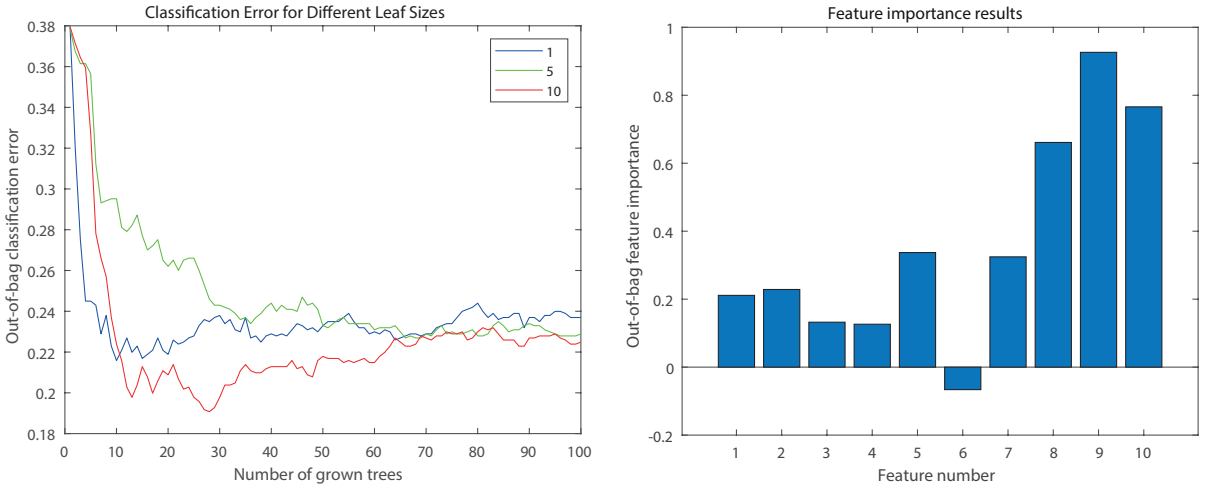


Рис. 2: График кучности данных об объеме инновационных товаров.

Fig. 2: Graph of the accuracy of data on the volume of innovative goods.

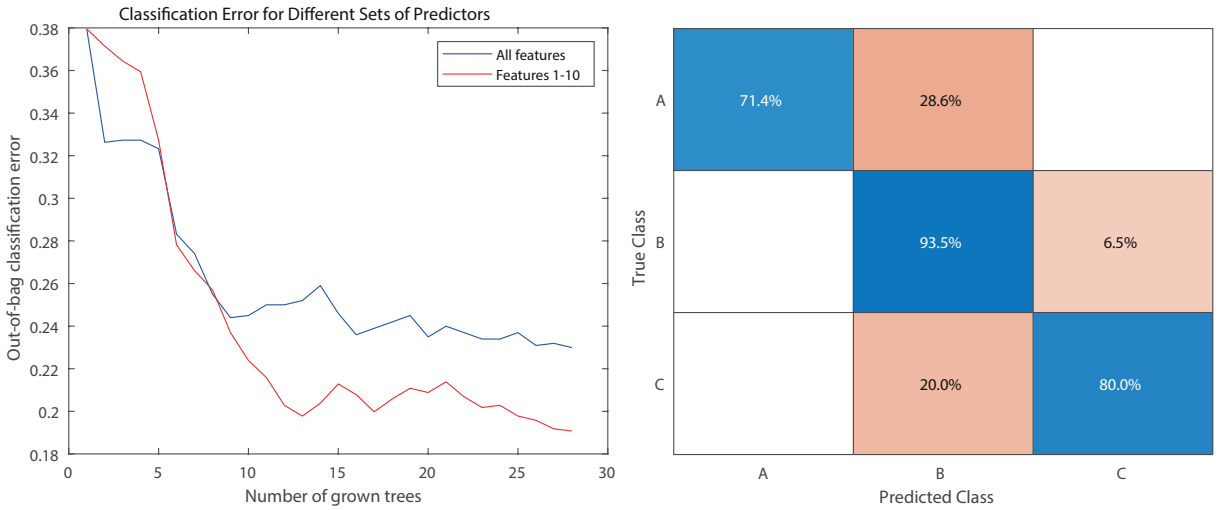
В 2022 г. регионы, у которых наблюдался фактический рейтинг A хотя бы для одной целевой функции, показаны в таблице 3. Среди полученного набора регионов для каждой входной переменной худшее и лучшее значения отмечены жирным шрифтом. Эти границы определяют сегменты входных переменных, которые являются плановыми для того, чтобы в будущем определить, будет ли регион иметь инновационный рейтинг A. При этом входные переменные представлены в ценах 2022 г., что также необходимо учитывать в дальнейшем.



a) b)

Рис. 3: Обучение случайных лесов с разной шириной листьев решающих деревьев (a); выбор наиболее важных переменных (b).

Fig. 3: Training random forests with different leaf widths of decision trees (a); selecting the most important variables (b).



a) b)

Рис. 4: Ошибка классификации для выбранных наиболее важных переменных (a); нормализованная матрица ошибок классификации (b).

Fig. 4: Classification error for selected most important variables (a); normalized classification error matrix (b).

Таблица 2: Предсказанные и фактические рейтинги регионов в 2022 г.

Table 2. Predicted and actual ratings of regions in 2022.

Номер наблюдения	**Предсказанный рейтинг**	**Вероятность рейтинга A**	**Вероятность рейтинга B**	**Вероятность рейтинга C**	**Фактический рейтинг**
997	'B'	0.066	0.754	0.18	'B'
998	'B'	0.046	0.552	0.403	'B'
...	...	...	...	...	...
1079	'C'	0.037	0.42	0.543	'C'

Таблица 3: Выбор регионов-лидеров по инновационным рейтингам и задание для их сегментов входных переменных.

Table 3: Selection of regions–leaders in innovative ratings and setting segments of input variables for them.

Регион	**Вход 1**	**Вход 2**	**Вход 3**	**Вход 4**	**Вход 5**	**Вход 6**	**Вход 7**	**Вход 8**
Московский	365433	22955	**50**	62.4	111899.7	14454.9	43173.8	104884.5
Москва	**3176924**	**460583**	57.1	**4591.4**	**2380209**	**96575**	**107000.4**	**278988.4**
С.-Петербург	401038	40194	62.4	664.3	271156.4	23307.5	27564	96864
Краснодарский	181561	13539	67.9	19.6	37893.4	3504	3571.5	**1155**
Татарстан	158605	14626	64.9	98.3	59219.7	3484.3	3729.7	18824.2
Нижегородский	146582	10984	66.1	33.3	36963.9	6669.4	11968.7	64834.2
Свердловский	218889	10627	68.1	53.8	52431.8	6513.4	3783.7	26123.1
Челябинский	**88323**	**5023**	**69.3**	**8.1**	**25657.2**	**637**	**1335.1**	21336.3
Новосибирский	142988	8600	65	40.5	51105	15509.8	6810.2	8455.8
Худшее значение	88323	5023	69.3	8.1	25657.2	637	1335.1	1155
Лучшее значение	3176924	460583	50	4591.4	2380209	96575	107000.4	278988.4

Регион	**Вход 9**	**Вход 10**	**Цель 1. Рейтинг. Факт**	**Цель 1. Рейтинг. Прогноз**	**Цель 2. Рейтинг. Факт**	**Цель 2. Рейтинг. Прогноз**	**Цель 3. Рейтинг. Факт**	**Цель 3. Рейтинг. Прогноз**
Московский	205693.1	**17461**	**'A'**	**'A'**	**'A'**	**'A'**	'B'	**'A'**
Москва	**722407.5**	15131	**'A'**	**'A'**	**'A'**	**'A'**	**'A'**	**'A'**
С.-Петербург	145751.9	13338	**'A'**	**'A'**	**'A'**	**'A'**	**'A'**	**'A'**
Краснодарский	36550.6	5436	'B'	'B'	**'A'**	'B'	'B'	'B'
Татарстан	258177	7264	**'A'**	**'A'**	**'A'**	'B'	'B'	'B'
Нижегородский	142304.8	8584	**'A'**	**'A'**	'B'	'B'	**'A'**	'B'
Свердловский	50644	14218	**'A'**	'B'	**'A'**	'B'	'B'	'B'
Челябинский	36929	7470	**'A'**	'B'	**'A'**	'B'	'C'	'B'
Новосибирский	**17648.8**	**3714**	'B'	'B'	'B'	'B'	**'A'**	'B'
Худшее значение	17648.8	3714						
Лучшее значение	722407.5	17461						

Заключение

1. Инновационный рейтинг регионов страны строится по трем целевым функциям: 1) объем инновационных товаров (всего) (млн руб.), 2) разработанные передовые производственные технологии (всего) (ед.), 3) сальдированный финансовый результат (информатизация и связь) (млн руб.). Таким образом, хотя рейтинг строится для отрасли РЭП, он также характеризует, как данная отрасль влияет на инновационное развитие и других отраслей промышленности – по целевым функциям 1 и 2. В этом заключается его преимущество.
2. Сам инновационный рейтинг подразумевает деление регионов на три класса: A – регионы-лидеры, B – регионы со средним уровнем инновационного развития, C – депрессивные регионы.
3. Построение и применение обозначенного инновационного рейтинга регионов подразумевает применение машинного обучения, а именно, решение задачи классификации с использованием технологии случайного леса, а далее верификация полученных рейтингов на данных нового периода наблюдения с целью определения по регионам-лидерам сегментов входных переменных модели. Эти сегменты являются плановыми показателями для того, чтобы в дальнейшем определить, будет ли регион иметь инновационный рейтинг A.
4. Точность предсказаний рейтингов регионов по объему инновационных товаров (всего) с помощью модели, обученной с использованием технологии случайного леса, на новых данных 2022 г. составляет 86.75%. Аналогично для разработанных передовых производственных технологий (всего) и сальдированного финансового результата (информатизация и связь) точность предсказаний рейтингов в 2022 г. составляет в обоих случаях 69.88%.

Полученные результаты могут быть полезны государственным структурам для определения, какие регионы следует поддерживать в их инновационном развитии отрасли радиоэлектронной промышленности и других отраслей промышленности.

Признательность: Исследование выполнено за счет гранта Российского научного фонда (проект № 24–28–00464).

Библиографический список

1. Доброва К.Б., Сахненко С.С. Предприятия радиоэлектронной промышленности в структуре высокотехнологичного сектора экономики // Экономика: вчера, сегодня, завтра. – 2022. – Т. 12. – №. 10–1. – С. 240–246. EDN: ZZLKNO.
2. Shi W.L. Industrial Electronics: Its Importance in the Manufacturing Industries // J Ind Electron Appl. – 2023. – Vol. 7 (1).
3. Ghosh D., Cabrera J. Enriched Random Forest for High Dimensional Genomic Data // ACM Journals: IEEEACM Transactions on Computational Biology and Bioinformatics. – 2022. – Vol. 19. – no. 5. – pp. 2817–2828. DOI: 10.1109/TCBB.2021.3089417.
4. Talagala T.S., Hyndman R.J., Athanasopoulos G. Meta–learning how to forecast time series // Journal of Forecasting. – 2023. – Vol. 42. – no. 6. – pp. 1476–1501. DOI: 10.1002/for.2963.
5. Sun Zh., Wang G., Li P., Wang H., Zhang M., Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees // Expert Systems with Applications. – 2024. – Vol. 237. – Part B. DOI: 10.1016/j.eswa.2023.121549.
6. Sipper M., Moore J.H. Conservation machine learning: a case study of random forests // Scientific Reports. – 2021. – Vol. 11. – pp. 3629. DOI: 10.1038/s41598-021-83247-4.

7. Tamilarasi S., Tsehay A.A., Subhashni R., Napa Komal K. A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model // International Journal of Electrical and Computer Engineering (IJECE). – 2022. – Vol. 12. – no. 2. – pp. 1831–1838. DOI: 10.11591/ijece.v12i2.pp1831-1838.
8. Abdulkareem N.M., Abdulazeez A.M. Machine Learning Classification Based on Radom Forest Algorithm: A Review // International Journal of Science and Business, IJSAB International. – 2021. – Vol. 5 (2). – pp. 128–142.
9. Loef B., Wong A., Janssen N.A.H., Strak M., Hoekstra J., Picavet H.S.J., Boshuizen H.H.C., Verschuren M.W.M., Herber G.–C.M. Using random forest to identify longitudinal predictors of health in a 30–year cohort study // Sci Rep. – 2022. – Vol. 12 (1). – pp. 10372. DOI: 10.1038/s41598-022-14632-w.
10. Wongvibulsin S., Wu K.C., Zeger S.L. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF–SLAM) data analysis // BMC Medical Research Methodology. – 2020. – Vol. 20(1). DOI: 10.1186/s12874-019-0863-0.
11. Legasa M.N., Manzanas R., Calvino A., Gutierrez J.M. A Posteriori Random Forests for Stochastic Downscaling of Precipitation by Predicting Probability Distributions // Water Resources Research. – 2022. – Vol. 58. – no. 4. DOI: 10.1029/2021WR030272.
12. Ooka T., Johno H., Nakamoto K., Yoda Y., Yokomichi H., Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large–scale health check-up data in Japan // BMJ Nutr Prev Health. – 2021. – Vol. 4 (1). – pp. 140–148. DOI: 10.1136/bmjnph-2020-000200.
13. Guo Y., Jia Ch., Wu Ch., Tu Y. Social Media Rumor Identification Based on Random Forest Classification and Feature Engineering: Case Study on Weibo Platform: Social Media Rumor Identification Based on Random Forest Classification // Proceedings of the 7th International Conference on Big Data and Computing (ICBDC '22). Association for Computing Machinery, New York, NY, USA, 2022. – pp. 109–118. DOI: 10.1145/3545801.3545817.
14. Klusowski J.M. Sharp Analysis of a Simple Model for Random Forests // Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research. – 2021. – Vol. 130. – pp. 757–765. URL: https://proceedings.mlr.press/v130/klusowski21b.html.
15. Hasan Dalfi M.A., Chaabouni S., Fakhfakh A. Breast Cancer Detection Using Random Forest Supported by Feature Selection // International Journal of Intelligent Systems and Applications in Engineering. – 2023. – Vol. 12. – no. 2s. – pp. 223–238. URL: https://ijisae.org/index.php/IJISAE/article/view/3575.

Vestnik of Samara University. Economics and Management. 2024, vol. 15, no. 4, pp. 187–201
doi: 10.18287/2542-0461-2024-15-4-187-201 ISSN: 2782-3008(online), ISSN 2542-0461(print)

Construction and application of innovative rating of regions using random forest technology

S. N. Yashin, N. I. Yashina, E. V. Koshelev

Lobachevsky State University, 23, Gagarin avenue, Nizhny Novgorod, 603022, Russia.

Abstract

The published article is devoted to the urgent problem of developing models for constructing and applying an innovative rating of the country's regions that have a radio-electronic industry (REP) branch. To overcome a number of shortcomings of known classical statistical approaches to improve the accuracy of forecasting the development of the REP industry and the adequacy of the ratings assigned to it. A version of the machine learning technology "random forest" has been developed and applied, increasing the accuracy of forecasting the development of the REP industry. The obtained ratings have been verified using data from a new observation period in order to determine the leading regions of the segments of the model's input variables.

Keywords: radio electronics industry; innovation rating;classification problem; random forest.

Received: Saturday 12th October, 2024 / Revised: Monday 11th November, 2024 / Accepted: Wednesday 11th December, 2024 / First online: Tuesday 28th January, 2025

Acknowledgments: The research was funded by the Russian Science Foundation (project No. 24–28–00464).

References

1. Dobrova K.B., Sakhnenko S.S. Enterprises of the radio-electronic industry in the structure of the high–tech sector of the economy // Economy: yesterday, today, tomorrow. – 2022. – Vol. 12. – No. 10–1. – pp. 240–246. EDN: ZZLKNO. (In Russ.)

Regional and Sectoral Economics (Research Article)

Please cite this article in press as:
Yashin S. N., Yashina N. I., Koshelev E. V. Construction and application of innovative rating of regions using random forest technology, *Vestnik Samarskogo Universiteta. Ekonomika i Upravlenie = Vestnik of Samara University. Economics and Management*, 2024, vol. 15, no. 4, pp. 187–201. doi: http://doi.org/10.18287/2542-0461-2024-15-4-187-201 (In Russian).

Authors' Details:
Sergey N. Yashin http://orcid.org/0000-0002-7182-2808
Doctor of Economics, Professor; Head of Department of Management and Public Administration; e-mail: jashinsn@yandex.ru

Nadezhda I. Yashina http://orcid.org/0000-0002-0630-7949
Doctor of Economics, Professor; Head of Department of Finance & Credit; e-mail: yashina@iee.unn.ru

Egor V. Koshelev http://orcid.org/0000-0001-5290-7913
Candidate of Economics, Associate Professor; Associate Professor of Department of Management and Public Administration; e-mail: ekoshelev@yandex.ru

2. Shi W.L. Industrial Electronics: Its Importance in the Manufacturing Industries // J. Ind. Electron Appl. – 2023. – Vol. 7 (1).
3. Ghosh D., Cabrera J. Enriched Random Forest for High Dimensional Genomic Data // ACM Journals: IEEEACM Transactions on Computational Biology and Bioinformatics. – 2022. – Vol. 19. – No. 5. – pp. 2817–2828. DOI: 10.1109/TCBB.2021.3089417.
4. Talagala T.S., Hyndman R.J., Athanasopoulos G. Meta–learning how to forecast time series // Journal of Forecasting. – 2023. – Vol. 42. – No. 6. – pp. 1476–1501. DOI: 10.1002/for.2963.
5. Sun Zh., Wang G., Li P., Wang H., Zhang M., Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees // Expert Systems with Applications. – 2024. – Vol. 237. – Part B. DOI: 10.1016/j.eswa.2023.121549.
6. Sipper M., Moore J.H. Conservation machine learning: a case study of random forests // Scientific Reports. – 2021. – Vol. 11. – pp. 3629. DOI: 10.1038/s41598-021-83247-4.
7. Tamilarasi S., Tsehay A.A., Subhashni R., Napa Komal K. A hybrid approach to medical decision–making: diagnosis of heart disease with machine-learning model // International Journal of Electrical and Computer Engineering (IJECE). – 2022. – Vol. 12. – No. 2. – pp. 1831–1838. DOI: 10.11591/ijece.v12i2.pp1831-1838.
8. Abdulkareem N.M., Abdulazeez A.M. Machine Learning Classification Based on Radom Forest Algorithm: A Review // International Journal of Science and Business, IJSAB International. – 2021. – Vol. 5 (2). – pp. 128–142.
9. Loef B., Wong A., Janssen N.A.H., Strak M., Hoekstra J., Picavet H.S.J., Boshuizen H.H.C., Verschuren M.W.M., Herber G.–C.M. Using random forest to identify longitudinal predictors of health in a 30–year cohort study // Sci Rep. – 2022. – Vol. 12 (1). – pp. 10372. DOI: 10.1038/s41598-022-14632-w.
10. Wongvibulsin S., Wu K.C., Zeger S.L. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF–SLAM) data analysis // BMC Medical Research Methodology. – 2020. – Vol. 20 (1). DOI: 10.1186/s12874-019-0863-0.
11. Legasa M.N., Manzanas R., Calvino A., Gutierrez J.M. A Posteriori Random Forests for Stochastic Downscaling of Precipitation by Predicting Probability Distributions // Water Resources Research. – 2022. – Vol. 58. – No. 4. DOI: 10.1029/2021WR030272.
12. Ooka T., Johno H., Nakamoto K., Yoda Y., Yokomichi H., Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large–scale health check–up data in Japan // BMJ Nutr Prev Health. – 2021. – Vol. 4 (1). – pp. 140–148. DOI: 10.1136/bmjnph-2020-000200.
13. Guo Y., Jia Ch., Wu Ch., Tu Y. Social Media Rumor Identification Based on Random Forest Classification and Feature Engineering: Case Study on Weibo Platform: Social Media Rumor Identification Based on Random Forest Classification // Proceedings of the 7th International Conference on Big Data and Computing (ICBDC '22). Association for Computing Machinery, New York, NY, USA, 2022. – pp. 109–118. DOI: 10.1145/3545801.3545817.
14. Klusowski J.M. Sharp Analysis of a Simple Model for Random Forests // Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research. – 2021. – Vol. 130. – pp. 757–765. URL: https://proceedings.mlr.press/v130/klusowski21b.html.
15. Hasan Dalfi M.A., Chaabouni S., Fakhfakh A. Breast Cancer Detection Using Random Forest Supported by Feature Selection // International Journal of Intelligent Systems and Applications in Engineering. – 2023. – Vol. 12. – No. 2s. – pp. 223–238. URL: https://ijisae.org/index.php/IJISAE/article/view/3575.