УДК 339.13.017
DOI 10.23968/1999-5571-2025-22-6-129-138

© *М. Б. Ласкин*, д-р экон. наук, гл. науч. сотрудник
(Санкт-Петербургский федеральный
исследовательский центр Российской академии наук,
Санкт-Петербург, Россия)
E-mail: laskin.m@iias.spb.su
© *В. Д. Крещенский*, аспирант
(Санкт-Петербургский государственный
университет, Санкт-Петербург, Россия)
E-mail: valerokr@gmail.com

© *M. B. Laskin*, Dr. Sci. Ec., Chief researcher
(Saint Petersburg Federal Research
Centre of the Russian Academy of Sciences,
St. Petersburg, Russia)
E-mail: laskin.m@iias.spb.su
© *V. D. Kreschensky*, post-graduate student
(Saint Petersburg State University,
St. Petersburg, Russia)
E-mail: valerokr@gmail.com

МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И АНСАМБЛЕВЫЕ АЛГОРИТМЫ РЕШАЮЩИХ ДЕРЕВЬЕВ В ЗАДАЧАХ МАССОВОЙ ОЦЕНКИ НЕДВИЖИМОГО ИМУЩЕСТВА

MULTIPLE LINEAR REGRESSION MODELS AND ENSEMBLE ALGORITHMS OF DECISION TREES IN PROBLEMS OF MASS REAL ESTATE EVALUATION

Рассматривается возможность применения ансамблевых алгоритмов решающих деревьев в задачах массовой оценки недвижимого имущества, включая кадастровую оценку. Результаты, полученные ансамблевыми методами, сравниваются с оценками, полученными по модели множественной линейной регрессии. Показано, что методы машинного обучения, основанные на решающих деревьях, могут быть точнее, чем регрессионные модели. Они могут применяться для верификации моделей, построенных традиционными оценочными техниками, а также обеспечивать повторяемость и воспроизводимость результатов.

Ключевые слова: массовая оценка недвижимого имущества, кадастровая оценка, методы машинного обучения, ансамблевые методы, решающие деревья.

The article considers the possibility of applying ensemble algorithms of decision trees in the problems of mass evaluation of real estate, including cadastral evaluation. The results obtained by ensemble methods are compared with the results of multiple linear regression models. It is shown that machine learning methods based on decision trees can be more accurate than regression models. They can be used for verification of models built by traditional evaluation techniques, as well as for providing repeatability and reproducibility of results.

Keywords: mass evaluation of real estate, cadastral evaluation, machine learning methods, ensemble methods, decision trees.

Введение

В задачах массовой оценки рыночной стоимости недвижимого имущества уже давно стали традиционными модели множественной линейной регрессии [1–3]. При построении таких моделей наряду с их очевидными преимуществами имеется и ряд сложностей, связанных с формализацией факторов и интерпретацией результатов. Одной из них является проблема ранговых факторов. Имеются в виду случаи, когда ранжирование уровней факторов затруднено или невозможно. При построении моделей множественной линейной регрессии проблема с ранговыми факторами может быть сведена к их трансформации в бинарные переменные [4, 5]. Такой подход дает хорошую интерпретируемость, но ведет к значительному увеличению размерности модели и, следовательно, требует большего количества объектов сравнения.

Ансамблевые алгоритмы, основанные на решающих деревьях, такой подготовки не требуют, они легко справляются как с би-

нарными, так и с ранговыми факторами. В оценочном сообществе существует устоявшееся мнение, что такие алгоритмы представляют собой некий «черный ящик», результаты которого воспроизвести невозможно, а схема построения алгоритма непонятна. Поэтому считается, что применение таких алгоритмов ведет к нарушениям стандартов оценки.

Цель настоящей статьи — показать, что алгоритмы, основанные на решающих деревьях, могут использоваться при оценке недвижимого имущества, они воспроизводимы, результаты могут быть не хуже, чем модели множественной линейной регрессии, и применяться при обработке больших объемов рыночных данных в задачах массовой оценки, включая определение кадастровой стоимости.

Следует отметить, что в научной литературе методы машинного обучения, основанные на решающих деревьях, давно рассматриваются как альтернатива множественной линейной регрессии, например, при моделировании пространственных распределений [6–8]. В частности, в работах [9, 10] решающие деревья рассматриваются как инструмент массовой оценки недвижимого имущества. При этом авторами [11–14] большое внимание в массовой оценке недвижимости уделяется и другим методам машинного обучения. В статье [15] на данных платформы Kaggle проводится сравнительный анализ результатов массовой оценки недвижимости, полученных методами регрессионного анализа и решающих деревьев.

В настоящей работе использованы материалы, предоставленные Санкт-Петербургским государственным бюджетным учреждением (СПб ГБУ) «Кадастровая оценка». Данные получены из открытых источников и предварительно обработаны для проведения кадастровой оценки встроенных объектов коммерческого назначения в Санкт-Петербурге.

Линейная регрессионная модель СПб ГБУ «Кадастровая оценка»

Рассматривается задача массовой оценки рыночной стоимости встроенных объектов коммерческого назначения в Санкт-Петербурге для определения кадастровой стоимости. Общий объем объектов сравнения — 8864. Ценообразующие факторы (далее ЦОФ): площадь объекта; влияние магистралей; влияние локальных центров; доступность общественного транспорта; этаж; класс качества здания, в котором расположен объект; тип объекта; тип входа.

Целевая переменная — скорректированная по первой группе корректировок (условия продажи, включая скидку на торг) цена за кв. м в тыс. руб. Часть факторов не требует преобразований, ранговые факторы преобразованы в бинарные переменные. Подбор модели множественной линейной регрессии дает модель, обладающую лучшими статистическими характеристиками:

- целевая переменная — логарифм скорректированной цены по 1 группе корректировок на условия продажи;
- ЦОФ — непрерывные и ранговые факторы, преобразованные в индикативные.

Модель получена применением библиотечной функции lm() статистического пакета *R*. Результат моделирования представлен в табл. 1.

Коэффициенты модели показаны в столбце Estimate. Как видно из данных таблицы, все коэффициенты статистически значимы (*p*-value *t*-критерия меньше 0,05); F-критерий Фишера указывает на адекватность модели (*p*-value меньше 0,05). Модель имеет высокое значение $R^2 = 0{,}8104$ и симметричные ошибки. Дополнительная проверка показывает, что ошибки модели не только симметричны, но и нормальны (рис. 1).

Результат проверки ошибок модели на нормальность тестом Колмогорова–Смирнова дает значение *p*-value = 0,08102, т. е. больше 0,05. Нулевой гипотезой в тесте

Таблица 1

Первичная модель множественной линейной регрессии

```
lm(formula = log(Скорректированная.цена.по.объявлению..руб..кв..м) ~
    Площадь..преобразованное.значение. + Влияние.магистралей.преобразованное.значение. +
        Влияние.локальных.центров.преобразованное.значение. + Доступность.общественного.транспорта.преобразованное.значение. +
        Выше первого+первый+цоколь+подвал+E+L+P+S+АДМ+БЦ+МФК+СТРИТ+ТК+С.улицы+Со.двора,
        data = dd)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91491 -0.16281 -0.00169  0.16006  0.88221

Coefficients:
                                                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                                     10.877037    0.017096  636.196  < 2e-16 ***
Площадь..преобразованное.значение.               0.320417    0.004073   78.677  < 2e-16 ***
Влияние.магистралей.преобразованное.значение.    0.066457    0.003109   21.374  < 2e-16 ***
Влияние.локальных.центров.преобразованное.значение.  0.187224  0.003697  50.645  < 2e-16 ***
Доступность.общественного.транспорта.преобразованное.значение. -0.090265  0.006141  -14.698  < 2e-16 ***
выше первого                                     0.714881    0.011953   59.810  < 2e-16 ***
первый                                           1.049792    0.008399  124.989  < 2e-16 ***
цоколь                                           0.499307    0.011486   43.470  < 2e-16 ***
подвал                                                NA          NA       NA       NA
E                                               -0.172563    0.006946  -24.844  < 2e-16 ***
L                                                0.336160    0.031536   10.660  < 2e-16 ***
P                                                0.278120    0.015921   17.469  < 2e-16 ***
S                                                     NA          NA       NA       NA
АДМ                                              0.125279    0.015966    7.846 < 4.78e-16 ***
БЦ                                               0.196999    0.016973   11.607  < 2e-16 ***
МФК                                              0.353299    0.022634   15.609  < 2e-16 ***
СТРИТ                                            0.248735    0.014169   17.555  < 2e-16 ***
ТК                                               0.166636    0.019597    8.503  < 2e-16 ***
ТОРГ                                                  NA          NA       NA       NA
С улицы                                          0.263935    0.005908   44.674  < 2e-16 ***
Со двора                                              NA          NA       NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2582 on 8847 degrees of freedom
Multiple R-squared:  0.8104,    Adjusted R-squared:  0.8101
F-statistic:  2363 on 16 and 8847 DF,  p-value: < 2.2e-16
```

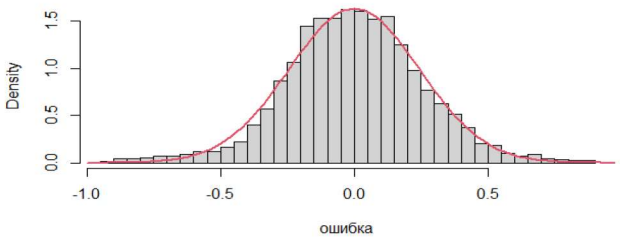

Рис. 1. Распределение ошибок модели. Красная линия — линия плотности нормального закона распределения с нулевым средним и стандартным отклонением RSE = 0,2582

Колмогорова–Смирнова является гипотеза о нормальности. Таким образом, при *p*-value = 0,08102 нет оснований отклонить данную гипотезу, качество модели вполне удовлетворительное. Исследование с применением функции regsubsets() библиотеки leaps статистического пакета *R* показывает, что улучшить эту модель не удается.

Исключенные из рассмотрения алгоритмом индикативные переменные «подвал» (один из рангов этажности), «S» (один из рангов класса качества здания), «ТОРГ» (один из рангов назначения помещения), «со двора» (один из рангов переменной «вход») исключены обоснованно — они уже учтены в свободном слагаемом. Вместо этих переменных из рассмотрения можно удалить любые другие индикативные переменные, по одной для каждой ранговой переменной. При удалении переменных «первый», «E», «СТРИТ», «с улицы» использование библиотечной функции lm() дает мо-

дель, полностью совпадающую с моделью, полученной ГБУ при кадастровой оценке. Распределение ошибок такое же (нормальное с нулевым средним и стандартным отклонением 0,25). Коэффициенты модели представлены в табл. 2.

В данном случае мы опускаем некоторые пояснения по построению модели множественной линейной регрессии, так как её построение оценочному сообществу хорошо известно. Эта модель построена для корректного сравнения с результатами оценок, полученных методами градиентного бустинга и алгоритмом «случайный лес».

Градиентный бустинг

Градиентный бустинг — метод последовательного улучшения дерева решений [16]. Обычно он дает лучшие результаты, чем алгоритм «Random Forest». Однако в данном случае результат будет чуть хуже, чем «Random Forest», но лучше, чем модель множественной линейной регрессии. Поскольку алгоритмы решающих деревьев построены на средних значениях, следует обратить вни-

Таблица 2

Модель множественной линейной регрессии СПб ГБУ «Кадастровая оценка»

```
lm(formula = log(Скорректированная.цена.по.объявлению..руб..кв..м) ~
    Площадь..преобразованное.значение. + Влияние.магистралей.преобразованное.значение. +
        Влияние.локальных.центров.преобразованное.значение. + Доступность.общественного.транспорта.преобразованное.значение. +
        Выше первого+цоколь+подвал+L+P+S+АДМ+БЦ+
        МФК+ТК+ТОРГ+Со.двора,data = dd)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91491 -0.16281 -0.00169  0.16006  0.88221

Coefficients:
                                                  Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                                     12.266936     0.007187  1706.941  < 2e-16 ***
Площадь..преобразованное.значение.               0.320417     0.004073    78.677  < 2e-16 ***
Влияние.магистралей.преобразованное.значение.    0.066457     0.003109    21.374  < 2e-16 ***
Влияние.локальных.центров.преобразованное.значение.  0.187224  0.003697    50.645  < 2e-16 ***
Доступность.общественного.транспорта.преобразованное.значение.  -0.090265  0.006141  -14.698  < 2e-16 ***
выше первого                                    -0.334911     0.009498   -35.261  < 2e-16 ***
цоколь                                          -0.550486     0.009582   -57.452  < 2e-16 ***
подвал                                          -1.049792     0.008399  -124.989  < 2e-16 ***
L                                                0.508724     0.031439    16.181  < 2e-16 ***
P                                                0.450683     0.015933    28.286  < 2e-16 ***
S                                                0.172563     0.006946    24.844  < 2e-16 ***
АДМ                                             -0.123457     0.009334   -13.226  < 2e-16 ***
БЦ                                              -0.051736     0.011841    -4.369  < 1.26e-05 ***
МФК                                              0.104563     0.018458     5.665  < 1.52e-08 ***
ТК                                              -0.082099     0.014923    -5.501  < 3.87e-08 ***
ТОРГ                                            -0.248735     0.014169   -17.555  < 2e-16 ***
Со двора                                        -0.263935     0.005908   -44.674  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2582 on 8847 degrees of freedom
Multiple R-squared:  0.8104,    Adjusted R-squared:  0.8101
F-statistic:  2363 on 16 and 8847 DF,  p-value: < 2.2e-16
```

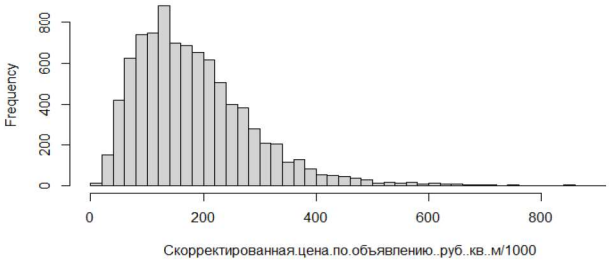

Рис. 2. Распределение цен объектов сравнения

мание на асимметрию распределения цен объектов сравнения (рис. 2).

Чаще всего распределения цен логарифмически нормальны [17–19], поэтому их логарифмирование приводит к изучению нормальных выборок. Мы уже получили лучшую модель множественной линейной регрессии для логарифма цены, что означает построение мультипликативной модели для цен. Поэтому ансамблевые алгоритмы решающих деревьев мы тоже применяем для логарифмов цен по полному набору ЦОФ. Рассмотрим результаты применения библиотечной функции gbm к нашим данным: ЦОФ — вещественные преобразованные переменные и ранговые переменные, целевая переменная — логарифм от скорректированных по 1-й группе цен за 1 кв. м. Бустинг — ансамблевый метод, поэтому визуализацию деревьев создать невозможно, но результаты поддаются анализу. Количество шагов — 2000. Множество объектов сравнения разбивалось (случайным образом) на обучающее множество (5000 объектов) и тестовое множество (3864 объекта). На рис. 3 показано соотношение между наблюдаемыми и предсказанными значениями логарифмов цен.

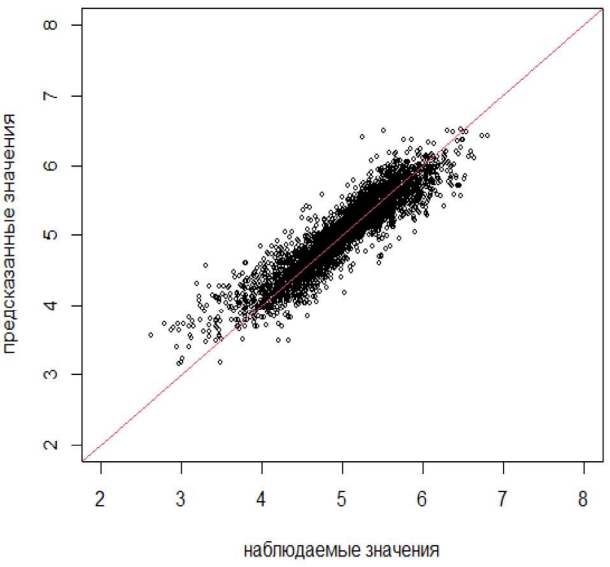

Рис. 3. Соотношение наблюдаемых значений логарифмов цен и предсказанных алгоритмом gbm

Ошибки модели gbm нормальны с нулевым средним и стандартным отклонением RSE=0,2303. Распределение ошибок показано на рис. 4.

Командная строка для запуска модели градиентного бустинга выглядит так:

gbm.res=gbm(log(Скорректированная.цена.по.объявлению..руб.. кв..м/1000) ~ Площадь..преобразованное.значение.+Влияние.магистралей...преобразован-ное.значение.+Влияние.локальных.центров...преобразованное.значение.+Доступность.общественного.транспорта...преобразованное.значение.+выше.пер-вого+первый+цоколь+подвал+E+L+P+S+АДМ+БЦ+МФК+СТРИТ+ТК+ТОРГ+с.улицы+со.двора,data=train,distribution=»gaussian»,n.trees=N,shrinkage=0.1,interaction.depth=4, bag.fraction=0.7,n.minobsinnode=5,cv.folds= 3,keep.data=F,verbose=F).

В правой части равенства среди прочих служебных записей перечислены все учитываемые в модели ЦОФ. Команда summary(gbm.res) дает возможность оценить значимость факторов (табл. 3).

Наибольшую важность имеют ранги «подвал», «первый» для этажности, площадь по-

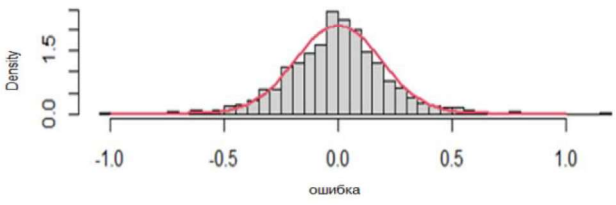

Рис. 4. Распределение ошибок модели gbm. Красная линия — линия плотности нормального закона распределения с нулевым средним и стандартным отклонением RSE = 0,2303; p-value теста Колмогорова–Смирнова на нормальность ошибок равно 0,1391

Таблица 3

Ранжирование ЦОФ факторов в порядке важности

var	rel.inf.
подвал	22.15638
первый	20.34905
Площадь..преобразованное.значение.	17.53124
Влияние.локальных.центров.преобразованное.значение.	16.80835
Влияние.магистралей.преобразованное.значение.	6.36611
Доступность.общественного.транспорта.преобразованное.значение.	4.89810
С улицы	4.10601
Е	1.72077
Со двора	1.61388
Р	0.91128
ТОРГ	0.76363
цоколь	0.66417
АДМ	0.42697
СТРИТ	0.42289
L	0.36651
выше.первого	0.27857
МФК	0.23818
S	0.21897
ТК	0.09235
БЦ	0.06658

мещения, влияние локальных центров. Следует отметить, что модель gbm лучше предсказывает значения на тестовом множестве, так как стандартное отклонение ошибок уменьшилось с RSE = 0,2582 до RSE = 0,2303 (см. рис. 1, 3).

Случайный лес (Random Forest)

Командная строка для запуска модели «Random Forest»:

bag.tt=randomForest(log(Скорректированная.цена.по.объявлению..руб..кв..м/1000)~Площадь..преобразованное.значение.+Влияние.магистралей...преобразованное.значение.+Влияние.локальных.центров...преобразованное.значение.+Доступность.общественного.транспорта...преобразованное.значение.+выше.первого+первый+цоколь+подвал+Е+L+Р+S+АДМ+БЦ+МФК+СТРИТ+ТК+ТОРГ+С.улицы+Со.двора,data = train, mtree = 20, ntree = 400, importance = TRUE).

Соотношение наблюдаемых значений логарифмов цен на тестовом множестве и предсказанных алгоритмом «Random Forest» показано на рис. 5.

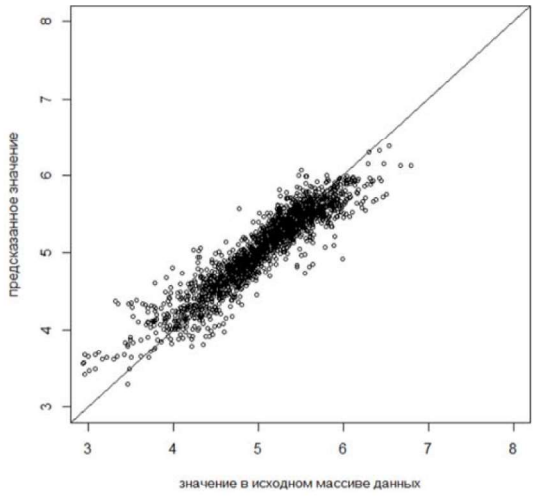

Рис. 5. Соотношение наблюдаемых значений логарифмов цен и предсказанных алгоритмом «Random Forest»

Распределение ошибок для модели «Random Forest» показано на рис. 6.

Ранжирование ЦОФ, участвующих в построении модели «Random Forest» в порядке важности, получаем применением

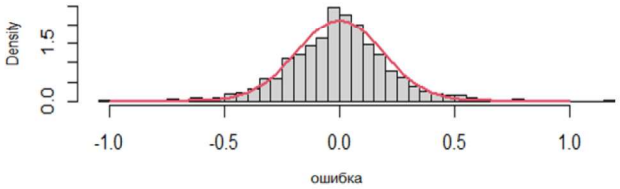

Рис. 6. Распределение ошибок модели «Random Forest». Красная линия — линия плотности нормального закона распределения с нулевым средним и стандартным отклонением RSE = 0,2185; p-value теста Колмогорова–Смирнова на нормальность ошибок равно 0,06608

библиотечной функции varImpPlot(bag.tt) (рис. 7).

Наибольшая значимость факторов в модели «Random Forest» такая же, как и в модели gbm: ранги «подвал», «первый» для этажности, площадь помещения, влияние локальных центров. Алгоритм «Random Forest» улучшает качество предсказаний по сравнению с моделью множественной линейной регрессии и моделью gbm, так как RSE еще раз снизилось до RSE = 0,2185.

Сравнение результатов

На рис. 8 показаны соотношения наблюдаемых значений логарифмов цен и предсказанных значений на тестовом множестве алгоритмами множественной линейной регрессии, gbm, «Random Forest».

Черным цветом показан результат по модели комбинированной множественной линейной регрессии, зеленым цветом — по модели градиентного бустинга, синим цветом — по модели случайного леса. Конечной целью исследования является сравнение не логарифмов цен, а собственно цен. С этой целью результаты, полученные во всех трех моделях, следует потенцировать. Сравнение наблюдаемых и предсказанных значений на тестовом множестве для всех трех алгоритмов показано на рис. 9.

Точность предсказаний, полученных по трем моделям, разная — множество точек, показанных синим цветом, более компактно, чем результаты двух других моделей. Ошибки во всех трех моделях нормальны с нулевым средним, но стандартные отклонения отличаются и равны:

- для модели множественной линейной регрессии RSE = 0,2582;
- для модели градиентного бустинга RSE = 0,2303;

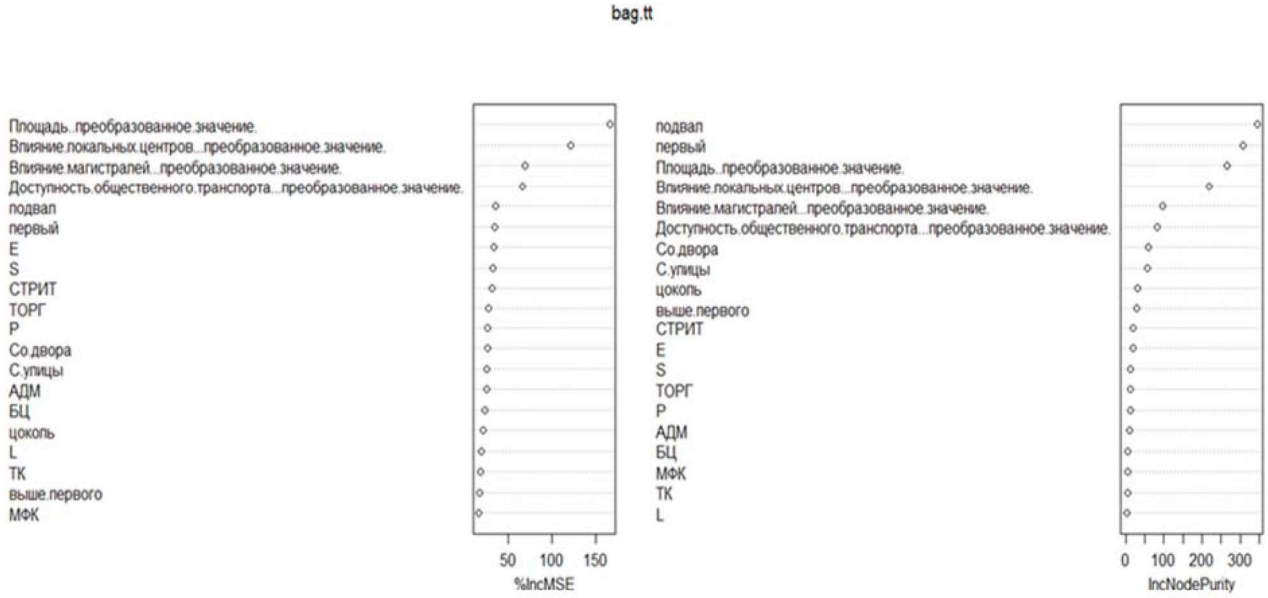

Рис. 7. Значимость факторов в модели Random Forest

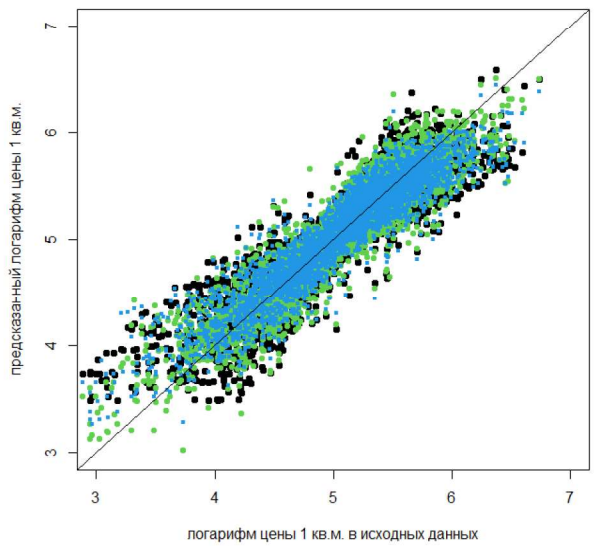

Рис. 8. Соотношение наблюдаемых значений логарифмов цен и предсказанных значений на тестовом множестве алгоритмами множественной линейной регрессии, gbm, «Random Forest»

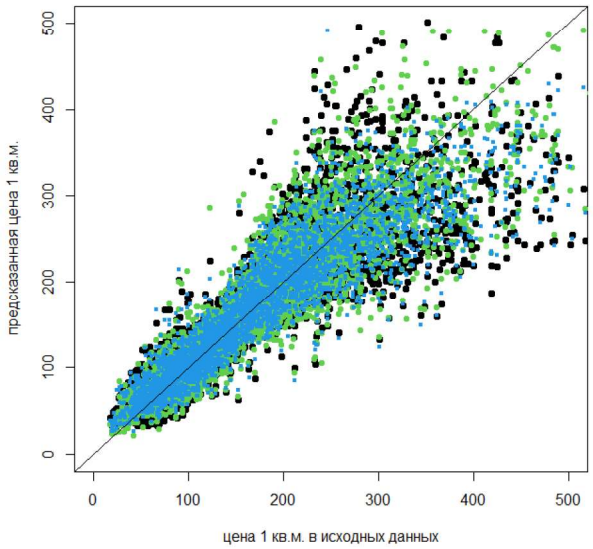

Рис. 9. Соотношение наблюдаемых значений цен и предсказанных значений на тестовом множестве алгоритмами множественной линейной регрессии, gbm, «Random Forest»

• для модели случайного леса («Random Forest») RSE = 0,2185.

Во всех случаях (см. рис. 9) имеется эффект увеличения отклонений по мере роста цены, т. е. в дорогом секторе. На рис. 8 можно видеть, что разброс увеличивается по мере как роста, так и уменьшения логарифма цены [20].

Таким образом, ансамблевые методы решающих деревьев могут давать результаты оценки лучше множественной линейной регрессии. Остается проблема соответствия таких методов расчета стандартам оценки и воспроизводимости результатов, так как считается, что при использовании той же информации и методов любой другой оценщик должен получить аналогичные результаты. В ансамблевых методах основным источником невоспроизводимости результатов может быть процедура случайного разбиения исходного датасета на обучающее и тестовое множества. Однако следует отметить, что в статистическом пакете R существует библиотечная функция *set.seed()*, с помощью которой устанавливается полная воспроизводимость работы генераторов случайных чисел и, как следствие, результатов оценки.

Выводы

Проведенное исследование позволяет сделать следующие выводы:

1. С алгоритмической точки зрения ансамблевые методы решающих деревьев могут применяться в задачах массовой оценки недвижимого имущества, в частности, при кадастровой оценке.

2. Результаты оценки такими методами могут быть даже точнее, чем оценка по модели множественной линейной регрессии.

3. Ансамблевые методы решающих деревьев могут, как минимум, использоваться для верификации традиционных оценочных техник.

4. Существуют библиотечные функции, методы и приемы, позволяющие добиться повторяемости и воспроизводимости результатов, полученных подобными методами машинного обучения.

Библиографический список

1. *Бурова И. В.* Использование регрессионного анализа в оценке стоимости объектов регионального

рынка недвижимости // Региональные проблемы преобразования экономики. 2020. № 2 (112). С. 39–45.

2. *Баринов Н. П.* Применение методов регрессионного анализа в задачах индивидуальной и массовой оценки объектов недвижимости // Вопросы оценки. 2022. № 1 (106). С. 34–46.

3. *Горобцова А. В.* Оценка рыночной стоимости квартир с помощью методов регрессионного анализа // Моделирование и анализ данных. 2019. Т. 9, № 2. С. 63–72.

4. *Анисимова И. Н., Баринов Н. П., Грибовский С. В.* Учет разнотипных ценообразующих факторов в многомерных моделях оценки недвижимости // Вопросы оценки. 2004. № 2. С. 2–15.

5. *Ласкин М. Б.* Множественная линейная регрессия и многомерные модельные распределения при оценке единых объектов недвижимости // Имущественные отношения в Российской Федерации. 2022. № 5 (248). С. 7–19; № 6 (249). С. 18–26.

6. *Ren X., Mi Z., Georgopoulos P. G.* Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States, Environ Int. 2020 Sep; 142:105827. DOI 10.1016/j.envint.2020.105827.

7. *Anmala J., Turuganti V.* Comparison of the performance of decision tree (DT) algorithms and extreme learning machine (ELM) model in the prediction of water quality of the Upper Green River watershed, Water Environ Res. 2021 Nov; 93(11):2360-2373. DOI 10.1002/wer.1642.

8. *Lalitha M., Dharumarajan S., Suputhra A., Kalaiselvi B., Hegde R., Reddy R. S., Prasad C. R. S, Harindranath C. S., Dwivedi B. S.* Spatial prediction of soil depth using environmental covariates by quantile regression forest model, Environ Monit Assess. 2021 Sep. 18;193 (10):660. DOI 10.1007/s10661-021-09348-9.

9. *Yilmazer S., Kocman S.* A mass appraisal assessment study using machine leaning based on multiple regression and random forest // Land Use Policy. 2020. Vol. 99. Article ID 104889. URL: https://doi.org/10.1016/j.landusepol.2020.104889

10. *Cordoba M., Carranza J.P., Piumetto M., Monzani F., Balzarini M.* A spatially based quantile regression forest model for mapping rural land values // Journal of Environmental Management. 2021. Vol. 289. No. 1. Article ID 112509. URL: https://doi.org/10.1016/j.jenvman.2021.112509

11. *Gu S., Kelly B., Xiu D.* Empirical asset pricing via machine learning // Chicago Booth Research Paper No. 18–04, 31st Australasian Finance and Banking Conference 2018, Yale ICF Working Paper No. 2018-09. 2019. URL: https://doi.org/10.2139/ssrn.3159577

12. *Kok N., Koponen E.-L., Martínez-Barbosa C. A.* Big data in real estate? From manual appraisal to automated valuation // The Journal of Portfolio Management. 2017. Vol. 43. No. 6. Pp. 202–211. URL: https://doi.org/10.3905/jpm.2017.43.6.202

13. *Steurer M., Hill R. J., Pfeifer N.* Metrics for evaluating the performance of machine learning based automated valuation models // Journal of Property Research. 2021. Vol. 38. No. 2. Pp. 99–129. URL: https://doi.org/10.1080/09599916.2020.1858937

14. *Kontrimas V., Verikas A.* The mass appraisal of the real estate by computational intelligence // Applied Soft Computing. 2011. Vol. 11. No. 1. Pp. 443–448. URL: https:// doi.org/10.1016/j.asoc.2009.12.003

15. *Yahan Fu,* A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models, 3rd International Conference on Applied Mathematics, Modeling Simulation and Automatic Control (AMMSAC 2024). 2024. Vol. 107. URL: https://doi.org/10.54097/vcy5n584

16. *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: data mining, inference, and prediction. New York: Springer Science & Business Media, 2009. 745 p.

17. *Aitchinson J., Brown J. A. C.* The Lognormal distribution with special references to its uses in economics. 1963, Cambridge: At the University Press.

18. *Ohnishi T., Mizuno T., Shimizu C., Watanabe T.* On the Evolution of the House Price Distribution. Columbia Business School. Center of Japanese Economy and Business, Working Paper Series. 2011. No 296. Pp. 1–20.

19. *Ласкин М. Б., Русаков О. В., Джаксумбаева О. И., Ивакина А. А.* Особенности формирования величины рыночной стоимости недвижимости при логарифмически нормальном распределении цен // Имущественные отношения в Российской Федерации. 2016. № 2 (173). С. 40–50.

20. *Баринов Н. П., Грибовский С. В., Зельдин М. А.* Точность результатов оценки и пределы ответственности оценщика // Имущественные отношения в Российской Федерации. 2009. № 9 (96). С. 43–55.

References

1. Burova I. V. *Ispol'zovanie regressionnogo analiza v otsenke stoimosti ob'ektov regional'nogo rynka nedvizhimosti* [Use of regression analysis in assessing the value of objects of the regional real estate market]. *Regional'nye problemy preobrazovaniya ekonomiki – Regional Problems of Economic Transformation*, 2020, no. 2 (112), pp. 39–45.

2. Barinov N. P. *Primenenie metodov regressionnogo analiza v zadachakh individual'noy i massovoy otsenki*

ob"ektov nedvizhimosti [Application of regression analysis methods in the problems of individual and mass evaluation of real estate objects. *Voprosy otsenki – Evaluation Issues,* 2022, no. 1 (106), no. 34–46.

3. Gorobtsova A. V. *Otsenka rynochnoy stoimosti kvartir s pomoshch'yu metodov regressionnogo analiza* [Assessment of the market value of apartments using regression analysis methods]. *Modelirovanie i analiz dannykh – Modeling and Data Analysis,* 2019, vol. 9, no. 2, pp. 63–72.

4. *Anisimova I. N., Barinov N. P., Gribovskiy S. V. Uchet raznotipnykh tsenoobrazuyushchikh faktorov v mnogomernykh modelyakh otsenki nedvizhimosti* [Accounting for various types of price-forming factors in multidimensional models of real estate valuation]. *Voprosy otsenki – Evaluation Issues,* 2004, no. 2, pp. 2–15.

5. Laskin M. B. *Mnozhestvennaya lineynaya regressiya i mnogomernye model'nye raspredeleniya pri otsenke edinykh ob"ektov nedvizhimosti* [Multiple linear regression and multivariate model distributions in the evaluation of single real estate objects]. *Imushchestvennye otnosheniya v Rossiyskoy Federatsii – Property Relations in the Russian Federation*, 2022, no. 5 (248), pp. 7–19, no. 6 (249), pp. 18–26.

6. *Ren X., Mi Z., Georgopoulos P. G.* Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States, *Environ Int.,* 2020, Sep., 142, 105827. DOI 10.1016/j.envint.2020.105827.

7. *Anmala J., Turuganti V.* Comparison of the performance of decision tree (DT) algorithms and extreme learning machine (ELM) model in the prediction of water quality of the Upper Green River watershed, *Water Environ Res.,* 2021, Nov., 93 (11), 2360-2373. DOI 10.1002/wer.1642.

8. *Lalitha M., Dharumarajan S., Suputhra A., Kalaiselvi B., Hegde R., Reddy R. S., Prasad C. R. S., Harindranath C. S., Dwivedi B. S.* Spatial prediction of soil depth using environmental covariates by quantile regression forest model. *Environ Monit Assess,* 2021, Sep 18, 193 (10), 660. Available at: https://link.springer.com/article/10.1007/s10661-021-09348-9

9. *Yilmazer S., Kocman S.* A mass appraisal assessment study using machine leaning based on multiple regression and random forest. *Land Use Policy*, 2020, vol. 99. Article ID 104889. Available at: https://doi.org/10.1016/j.landusepol.2020.104889

10. *Cordoba M., Carranza J. P., Piumetto M., Monzani F., Balzarini M.* A spatially based quantile regression forest model for mapping rural land values. *Journal of Environmental Management,* 2021, vol. 289, no. 1. Article ID 112509. Available at: https://doi.org/10.1016/j.jenvman.2021.112509

11. *Gu S., Kelly B., Xiu D.* Empirical asset pricing via machine learning. Chicago Booth Research Paper No. 18–04. *Proceedings of the 31st Australasian Finance and Banking Conference*, 2018, Yale ICF Working Paper no. 2018-09, 2019. Available at: https://doi.org/10.2139/ssrn.3159577

12. *Kok N., Koponen E.-L., Martínez-Barbosa C.A.* Big data in real estate? From manual appraisal to automated valuation. The Journal of Portfolio Management, 2017, vol. 43, no. 6, pp. 202–211. Available at: https://doi.org/10.3905/jpm.2017.43.6.202

13. *Steurer M., Hill R. J., Pfeifer N.* Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research,* 2021, vol. 38, no. 2, pp. 99–129. Available at: https://doi.org/10.1080/09599916.2020.1858937

14. *Kontrimas V., Verikas A.* The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 2011, vol. 11, no. 1, pp. 443–448. Available at: https://doi.org/10.1016/j.asoc.2009.12.003

15. *Yahan Fu.* A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models, Proceedings of the 3-rd International Conference on Applied Mathematics, Modeling Simulation and Automatic Control (AMMSAC 2024), vol. 107 (2024). Available at: https://doi.org/10.54097/vcy5n584

16. *Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction.* New York: Springer Science & Business Media, 2009, 745 p.

17. *Aitchinson J., Brown J. A. C. The Lognormal distribution with special references to its uses in economics.* 1963, Cambridge, University Press Publ.

18. *Ohnishi T., Mizuno T., Shimizu C., Watanabe T.* On the Evolution of the House Price Distribution. Columbia Business School. Center of Japanese Economy and Business, Working Paper Series, 2011, no 296, pp. 1–20.

19. Laskin M. B., Rusakov O. V., Dzhaksumbaeva O. I., Ivakina A. A. *Osobennosti formirovaniya velichiny rynochnoy stoimosti nedvizhimosti pri logarifmicheski normal'nom raspredelenii tsen* [Features of the formation of the market value of real estate with a logarithmic normal distribution of prices]. *Imushchestvennye otnosheniya v Rossiyskoy Federatsii – Property relations in the Russian Federation*, 2016, no. 2 (173), pp. 40–50.

20. Barinov N. P., Gribovskiy S. V., Zel'din M. A. *Tochnost' rezul'tatov otsenki i predely otvetstvennosti otsenshchika* [Accuracy of evaluation results and limits of responsibility of the appraiser]. *Imushchestvennye otnosheniya v Rossiyskoy Federatsii – Property relations in the Russian Federation*, 2009, no. 9 (96), pp. 43–55.