УДК 528.94:332.2
DOI 10.33764/2411-1759-2023-28-5-103-112

Использование предиктивной геоинформационной модели для оценки рыночной стоимости земельных участков в экспертной системе

А. В. Курлов[1]

[1] Научно-исследовательская лаборатория «Лаборатория городских технологий и пространственного развития», г. Москва, Российская Федерация
* e-mail: kurlov-av@yandex.ru

Аннотация. При решении задачи прогнозирования рыночной стоимости существует проблема, связанная с набором данных по земельным участкам. Для определения реальной рыночной стоимости была выдвинута гипотеза о том, что модель для предсказания кадастровой стоимости участка будет эффективна и для предсказания увеличения рыночной стоимости земельного участка. Исследование направлено на определение рыночной стоимости земельных участков с учетом пространственного фактора. В статье анализируются данные о рыночной стоимости участков на основе зарегистрированных сделок купли-продажи. Результаты исследования показали сильное влияние местоположения земельных участков относительно социальной инфраструктуры, федеральных трасс, крупных населенных пунктов и водоемов. На основе этого разработана математическая модель, использующая ансамблевый самооптимизирующийся метод, для определения стоимости квадратного метра земельного участка. Принцип работы будущего программного продукта заключается в том, что в специальном плагине пользователь может выбрать земельный участок на карте или задать его координаты и получить на выходе кластер, гипотетическую стоимость земельного участка и прибыль за некоторое заданное количество лет.

Ключевые слова: рыночная стоимость, кадастровая стоимость, анализ, прогнозирование, регрессионная модель, экспертная модель, CatBoost, NgBoost, XGBoost

Введение

Определение реальной рыночной стоимости не только является важной задачей с экономической и финансовой точки зрения, но и позволяет оценивать развитие территории, населенных пунктов и их районов. Данные о рыночной стоимости участка могут формироваться несколькими способами, один из них – это данные о зарегистрированных сделках купли-продажи. Основная проблема заключается в том, что невозможно по одному земельному участку таким способом получить ежегодные данные о стоимости, поэтому работа с такой информацией затруднена и построение модели без предварительной обработки данных невозможно.

Методы и материалы

Для построения прототипа и тестирования моделей можно использовать данные о кадастровой стоимости земельных участков, наиболее релевантные модели протестировать на наиболее заполненных участках в данных по рыночной стоимости. Ежегодно данные о новых участках и их кадастровой стоимости обновляются Единым государственным реестром недвижимости, поэтому наличие пропущенных данных о стоимости в разные годы маловероятно, однако если они имеются, то их число незначительно и не оказывает критического влияния на общую точность данных.

Проведем исследование данных по кадастровой и рыночной стоимости земельных участков. Процесс обработки и подготовки данных для построения модели можно разделить на несколько этапов:

– анализ признаков и полноты данных;

– выделение целевой переменной;

– исследование важности признаков, очистка данных.

В данных имеется информация о стоимости земельных участков, предназначенных для индивидуального жилищного строительства (ИЖС), их площади, координатах, расстоянии до объектов инфраструктуры, рек, населенных пунктов и др.

В данных по кадастровой стоимости имеется 17 145 участков, по каждому из которых описано более 30 признаков. Имеются объекты, в которых пропущена информация, например, кадастровая стоимость в разные годы, отсутствует наименование муниципалитета, нет атрибутов участка. Отбросим признаки, в которых отсутствует более 30 % информации, таким образом, набор данных значительно сократился, количество признаков стало равным 15. В качестве признаков сохранились кадастровая стоимость земельных участков, расстояние до трассы М11, школы, детского сада, поликлиники, р. Волги, Твери, Москвы.

В ходе исследования было выяснено, что пространственные факторы, такие как расстояние до объектов инфраструктуры, водоемов, крупных населенных пунктов, существенно влияют на конечную стоимость земельных участков для ИЖС. Таким образом, считаем сокращение размерности набора данных обоснованным не только с математической точки зрения, но и с практической.

Данные обладали разбросом значений стоимости участка от 1 рубля до 1,8 млрд рублей (рис. 1), отнести получившееся распределение к какому-либо типу затруднительно. Проведем нормализацию данных. Слишком высоких и низких значений стоимости участков не так много, поэтому считаем их промахами, т. е. результатами наблюдений, которые резко отличаются от остальных [1]. При статистическом анализе данных такие участки отбрасывают в соответствии с межквартильным анализом.

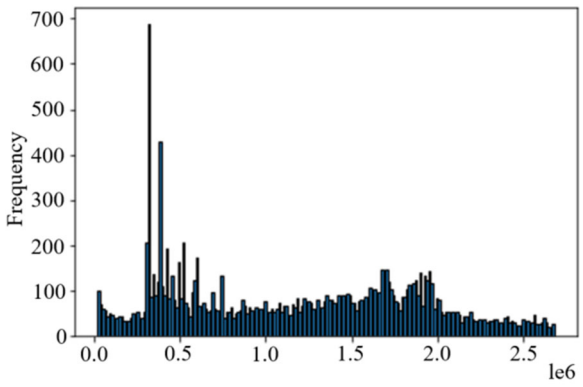

Рис. 1. Распределение стоимости земельных участков ИЖС до нормализации данных

Проведем анализ межквартильного диапазона (разница между первым и третьем перцентилями) по значениям кадастровой стоимости cad_cost. Это позволит оценить разброс значений в наборе значений кадастровой стоимости. Преимуществом использования межквартильного диапазона является возможность рассмотреть разброс средних значений в выборке без учета выбросов. Возьмем стандартные для нормального распределения значения перцентилей ($Q1$ = 75 %, $Q3$ = 25 %) и построим график распределения «ящик с усами» (рис. 2).

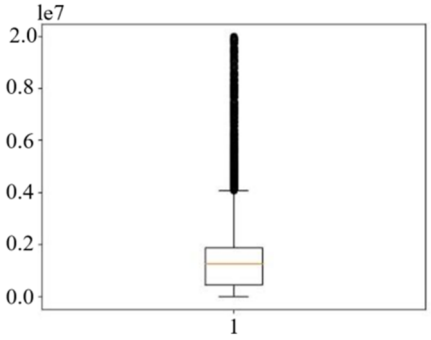

Рис. 2. Распределение кадастровой стоимости до обработки

Медианное значение равняется 1 299 000 руб. Значение верхнего квантиля составляет 1 900 000 руб. Значение нижнего квантиля равно 464 000 руб. Вычислим межквартильный размах IQR:

$$IOR = Q1 - Q3 = 1\,436\,000 \text{ руб.}$$

В представленных данных минимальное значение стоимости земельного участка недвижимости составляет 1 рубль, максимальное – 1 800 000 000 руб. Заметим, что жирной линией отмечена граница выбросов.

Подберем оптимальные значение доли межквартильного диапазона. Стандартное значение данного коэффициента для нормального распределения составляет 1,5. С учетом распределения кадастровой стоимости, возьмем долю верхнего диапазона, равную 1,8, нижнюю долю, равную 0,3. Теперь отбросим значения, которые выходят за пределы интервала, назовем их выбросами. Вы-

бросами считаются все значения кадастровой стоимости, не удовлетворяющие условию «Кадастровая стоимость меньше, чем ($Q1-1{,}8*IQR$)» или «Кадастровая стоимость больше, чем ($Q3-0{,}3*IQR$)».

Построим график распределения (рис. 3). Визуально распределение стало ближе к нормальному, несмотря на смещение. Минимальная стоимость земельного участка составляет 24 000 руб., максимальное – 2 499 000 руб. (рис. 4). Жирная линия, отражающая выбросы, теперь отсутствует. Обновленное количество исследуемых земельных участков составило 15 578 земельных участков. Таким образом, отброшено около 9 % от изначального датасета.

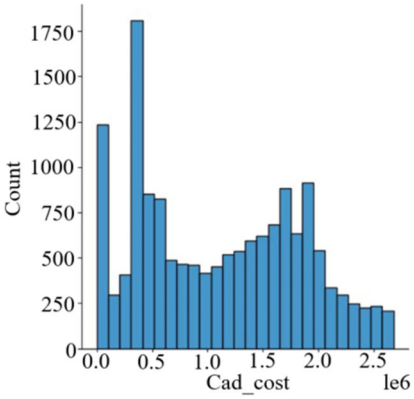

Рис. 3. Распределение стоимости земельных участков ИЖС после нормализации данных

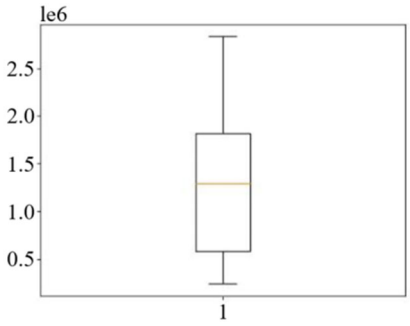

Рис. 4. Распределение кадастровой стоимости после обработки

Найдем медианное значение стоимости земельных участков в каждый год во временном интервале с 2015 по 2021 г., построим зависимость, которая имеет практически линейный характер (рис. 5).

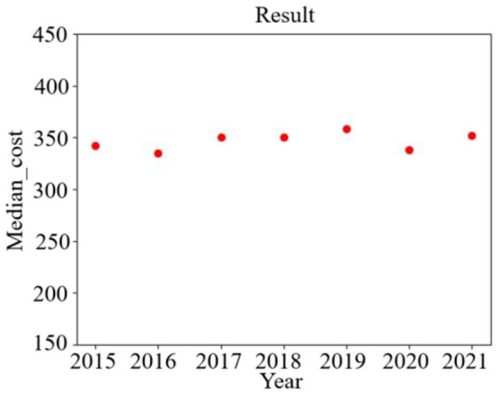

Рис. 5. Медианная стоимость земельных участков во временном интервале с 2015 по 2021 г.

Построение прогнозной модели для определения кадастровой стоимости на основе классических регрессионных подходов

Для осуществления предсказания рыночной стоимости выдвинута гипотеза, что модель для предсказания кадастровой стоимости участка будет эффективна и для предсказания увеличения рыночной стоимости земельного участка. Задача относится к классу регрессионных задач, поэтому для обучения моделей воспользуемся методами регрессионного анализа.

Регрессионные модели представляют собой обширный класс моделей, которые позволяют описать и выявить закономерность между зависимой и независимыми переменными [2].

В качестве целевой (зависимой) переменной рассмотрим кадастровую стоимость земельного участка *cad_cost*, признаками (независимыми переменными) выступают пространственные факторы.

Рассмотрим базовые методы регрессионного анализа, а именно линейные модели: линейная регрессия, полиномиальная регрессия, сингулярное разложение, регрессия Лассо, гребневая регрессия, метод опорных векторов.

Линейная регрессия [3] является методом, позволяющим подбирать весовые коэффициенты *wi*, *bi* для построения линейной зависимости между целевой переменной y и признаками *xi*:

$$y = \sum w_i x_i + b_i.$$

Линейная регрессия использует совокупность независимых переменных для объяснения или прогнозирования результата. Подбор весовых коэффициентов основан на минимизации суммы квадратов между линейным приближением и истинным значением целевой переменной.

Зависимость целевой переменной от признаков может быть намного сложнее, чем линейная, и в таких случаях можно использовать алгоритм полиномиальной регрессии, которая позволяет построить нелинейную зависимость в соответствии с полиномом *n* порядка. Полиномиальная регрессия – это алгоритм машинного обучения, который используется для обучения линейной модели на нелинейных данных [3]. Общий вид уравнения для модели полиномиальной регрессии выглядит следующим образом:

$$y = b_0 + \sum b_i x_{i-1}{}^{i}.$$

Деревья решений [3] используются для подбора синусоидальной кривой с добавлением зашумленных наблюдений. Такой алгоритм основан на исследовании локальных линейных регрессий, аппроксимирующих синусоиду. Максимальная глубина дерева может быть установлена слишком большой, в этом случае деревья решений изучают слишком мелкие детали обучающих данных и учатся на шуме, в результате чего может возникать переобучение модели.

Метод регрессии Лассо [3] использует метод наименьших квадратов для определения коэффициентов модели:

$$y = X\beta + \varepsilon;$$

$$loss(\beta, \lambda) = \|y - X\beta\|_2^2 + \lambda \sum b_i,$$

где λ – параметр регуляризации метода; β и ε – коэффициенты модели.

Гребневая регрессия, являющаяся родственной регрессии Лассо, основана на минимизации квадрата абсолютной суммы коэффициентов модели.

Метод опорных векторов [3] заключается в построении уравнения пространства векторов, в котором квадратичные потери являются минимальными. Если вектор попадает в пространство, построенное на обучающей выборке, потери считаются равными нулю. В таком случае говорят о 100-процентного попадании в истинное значение. Если этого не происходит, то оценивается отклонение фактического и прогнозируемого значений. Для построения пространства используется некоторое ядро, которое представляет собой полиномиальную, линейную, радиально-базисную, сигмоидную функции или функцию Фурье.

Сингулярное разложение *SVD*, или разложение Шмидта, – это декомпозиция вещественной матрицы с целью ее приведения к каноническому виду. Разложение Шмидта показывает геометрическую структуру матрицы и позволяет наглядно представить имеющиеся данные [4]. *SVD* позволяет вычислять обратные и псевдообратные матрицы большого размера, что делает его полезным инструментом при решении задач регрессионного анализа.

Обученные алгоритмы машинного обучения на основе регрессионных моделей требуют оценки релевантности их работы. Для оценки точности моделей используют метрики качества – функции, по которым определяется качество обученной модели машинного обучения (как на всем датасете в целом, так и на отдельных моделях), но не происходит непосредственной оптимизации. Используются метрики и для контроля переобучения, для сравнения различных моделей.

Понятными и интерпретируемыми метриками являются средняя абсолютная ошибка (*MAE*, от англ. *mean squared error*) и относительная ошибка (*MAPE*, от англ. *mean absolute percentage error*) [5]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|a(x_i) - y_i\right|;$$

$$MAPE = \frac{1}{n} \frac{\sum_{i=1}^{n} |a(x_i) - y_i|}{\sum_{i=1}^{n} |y_i|}.$$

Здесь *yi* – истинное значение целевой переменной; *a*(*xi*) – значение целевой переменной, спрогнозированное моделью. Другими словами, *MAE* показывает среднее отклонение предсказанного значения от истинного, *MAPE* выражает ошибку прогноза в процентах. Для оценки точности регрессионного прогноза необходимо иметь значения с уже известным результатом стоимости земельных участков, чтобы сравнить с предсказанным.

Поэтому изначально разбиваем данные о кадастровой стоимости случайным образом на две выборки: обучающую, на которой будут строиться регрессионные модели (она составляет 80 % от предобработанного набора данных), тестовую, на которой будем оценивать релевантность регрессионных моделей (составляет 20 % от предобработанного набора данных). На обучающей выборке строятся модели на основе алгоритмов линейной, полиномиальной регрессии, дерева решений, метода Лассо, гребневой регрессии, сингулярного разложения. Затем на тестовой выборке происходит оценка точности работы алгоритмов.

Для всех вышеперечисленных алгоритмов метрика *MAE* составила около 1 000 000 руб., что нельзя считать приемлемым результатом, поскольку для некоторых земельных участков метрика *MAE* превышает стоимость самих участков на порядок.

Построение прогнозной модели для определения кадастровой стоимости на основе ансамблевых регрессионных подходов

Рассмотрим семейство ансамблевых моделей, позволяющих не только строить регрессионные зависимости, но и осуществлять оптимизацию в процессе обучения. Ансамбли (комбинации) математических подходов, в основе которых лежат бустинговые алгоритмы и деревья решений, показали хороший результат при прогнозировании финансовых процессов [6].

Бустинговые модели [7] – это обширный класс моделей машинного обучения, который представляет собой ансамблевые мета-алгоритмы, иными словами композицию алгоритмов с возможностью их оптимизации в процессе обучения. Бустинговые модели основаны на построении в ходе итерационного процесса функций, причем на каждом шаге модель обучается с использованием данных об ошибках предыдущих. С каждой итерацией изначально слабая функциональная зависимость становится сильнее, на выходе получается так называемая сильная функция.

В качестве тестового воспользуемся алгоритмом *NgBoost*. В основе алгоритма лежит построение модели по параметрическому распределению вероятностей. Алгоритм является модульным: состоит из нескольких взаимосвязанных блоков. В одном блоке осуществляется обучение модели регрессионного дерева решений, в другом оценивается параметрическое распределение вероятностей, в третьем применяется метрика для оценки точности и оптимизации модели, именно в третьем модуле выбираются наилучшие параметры модели.

Для оптимизации работы модели воспользуемся в качестве функции потерь метрикой *MAE*. Ее конкуренты, метрики *RMSE* (корень из средней квадратичной ошибки) и *MAPE*, сильно чувствительны к грубым выбросам. В качестве примера, при обучении сравним показатели метрик *MAPE, RMSE* и *MAE* для прогнозных значений на тестовой выборке (табл. 1). Наилучший результат получился в случае применения метрики в качестве функции потерь при оптимизации ансамблевого алгоритма *NgBoost*.

В качестве обучающей и тестовой выборки использовали то же разбиение, что и для классических регрессионных моделей: 80 % исходных данных – тестовая выборка; 20 % данных – обучающая. После выбора наиболее релевантных параметров модели значение метрики средней

абсолютной ошибки значительно уменьшилось и составило *MAE* = 152 197 руб. Ошибка значительно снизилась в сравнении с классическими подходами, обучим другие родственные алгоритмы: *XGBoost, LightGBM, Catboost*.

Таблица 1

Результаты применения метрик в качестве функций потерь

RMSE в качестве функции потерь	*MAPE* в качестве функции потерь	*MAE* в качестве функции потерь
$\sqrt{\frac{\sum_{i=1}^{N}(a_i - t_i)^2 w_i}{\sum_{i=1}^{N} w_i}}$	$\frac{\sum_{i=1}^{N} w_i \frac{\lvert a_i - t_i \rvert}{Max(1, \lvert t_i \rvert)}}{\sum_{i=1}^{N} w_i}$	$\frac{\sum_{i=1}^{N} w_i \lvert a_i - t_i \rvert}{\sum_{i=1}^{N} w_i}$
R2: 0.13 RMSE: 12555873 MAE: 1198353.7	R2: -0.01 RMSE: 13563525.4 MAE: 1977656.81	R2: 0.32 RMSE: 11083893 MAE: 810144.88

XGBoost (полное название *Extreme Gradient Boosting*) представляет собой масштабируемую распределенную модель машинного обучения с градиентным бустингом дерева решений. Она обеспечивает параллельный бустинг деревьев, хорошо решает задачи регрессии, классификации и ранжирования.

Машинное обучение с учителем использует алгоритмы для обучения модели поиску шаблонов в наборе данных с метками и функциями, а затем использует обученную модель для прогнозирования меток для функций нового набора данных. Деревья решений создают модель, которая предсказывает метку, оценивая дерево вопросов по признакам «если-то-иначе», «истина/ложь» и минимальное количество вопросов, необходимых для оценки вероятности принятия правильного решения. Деревья решений можно использовать для решения задач регрессии, чтобы предсказать непрерывное числовое значение.

Градиентный бустинг деревьев решений – это алгоритм обучения ансамбля деревьев решений, аналогичный случайному лесу, для классификации и регрессии. Алгоритмы ансамблевого обучения объединяют несколько алгоритмов машинного обучения для получения лучшей модели [8–12].

И случайный лес, и градиентный бустинг деревьев решений строят модель, состоящую из нескольких деревьев решений. Разница в том, как деревья построены и объединены. Случайный лес использует метод, называемый бэггингом, для параллельного построения полных деревьев решений из случайных выборок набора данных. Окончательный прогноз представляет собой среднее значение всех прогнозов дерева решений.

Термин «градиентного бустинга» происходит от идеи улучшения одной слабой модели путем объединения ее с рядом других слабых моделей для создания сильной модели. Градиентный бустинг – это процесс объединения слабых моделей с применением градиентного спуска к целевой функции [13].

XGBoost – это масштабируемая и высокоточная реализация градиентного бустинга, которая расширяет пределы вычислительной мощности алгоритмов градиентного бустинга деревьев решений. С *XGBoost* деревья строятся параллельно, а не последовательно, как при градиентном бустинге деревьев решений. Он работает последовательно, уровень за уровнем сканируя значения градиента и используя эти частичные суммы для оценки качества разбиений при каждом возможном разбиении в обучающем наборе.

CatBoost – ансамблевый алгоритм градиентного бустинга и деревьев решений. Метод основан на построении «небрежных» деревьев решений, далее с помощью оптимизации осуществляется автоматизированный выбор оптимального дерева. Особенностью *Catboost* является его способность автоматически обрабаты-

вать категориальные признаки, что упрощает и ускоряет процесс подготовки данных. Кроме того, он использует методы ранней остановки и случайного сэмплирования, благодаря чему позволяет избежать переобучения модели и повысить ее обобщающую способность [14, 15].

Среди всех описанных выше алгоритмов лучший результат показала модель *CatBoost* с минимальной метрикой *MAE* = 63 474 руб. при среднем значении стоимости участка 670 000 руб. Проведем оптимизацию модели и подберем ее параметры (табл. 2).

Таблица 2

Параметры алгоритма *CatBoost*

Наименование и значение параметра	Описание
loss_function='MAE	В качестве функции потерь используем метрику *MAE*, показавшую наилучшую сходимость результата прогноза с истинным значением кадастровой стоимости
eval_metric='MAE'	Метрика для оценки точности модели – *MAE*
iterations=25000	Число эпох обучения составило 25 000
early_stopping_rounds=1000	Число итераций – 1 000. Если на валидации алгоритм получил результат хуже, чем ранее, модель перестает обучаться. Это говорит о том, что лучший результат определен
depth=8	Глубина дерева – 8 (количество возможных ответвлений вглубь дерева)
verbose=1000	Количество итераций, через которые визуализируются валидационные метрики в процессе обучения. Параметр позволяет визуально контролировать процесс обучения модели
grow_policy='Lossguide'	Правило построения листьев дерева решений
max_leaves=64	Максимальное количество листьев в дереве
l2_leaf_reg=3	Коэффициент регуляризации
min_data_in_leaf=100	Минимальное количество данных в листьях для построения дерева

Корректировка целевой переменной

В процентном соотношении ошибка работы алгоритма составляет целых 9 %, а само значение отклонения от истинного значения для некоторых участков сопоставимо с их стоимостью. Так, для земельных участков ИЖС стоимостью 250 000 руб. погрешность составляет целых 25 %, что едва ли можно считать объективной оценкой.

Скорректируем целевую переменную. При заключении договоров купли-продажи, помимо стоимости самих участков, указывается и цена за квадратный метр. Земельные участки, представленные в наборе данных о кадастровой стоимости, имеют разную площадь. Участки одной и той же площади могут иметь одинаковую стоимость, находясь в разных населенных пунктах. Поэтому корректнее говорить о стоимости участка за квадратный метр.

Сгенерируем новую переменную: стоимость квадратного метра земельного участка ИЖС *sqr_cost* как частное стоимости земельного участка и его площади. Проведем обучение классических и ансамблевых бустинговых регрессионных моделей. Результаты обучения приведены в табл. 3 при средней стоимости земельного участка ИЖС за квадратный метр 923 руб.

Таблица 3

Результаты обучения регрессионных моделей для прогнозирования кадастровой стоимости земельных участков

Прогнозная модель	Значение метрики *MAE*, руб.
Дерево решений	550
SVM	678
NgBoost	100,6
CatBoost	47,3

Результаты

Наиболее релевантный результат получился с помощью алгоритма CatBoost, как это было и ранее. Ошибка составила всего лишь 5 %. Принимаем в качестве алгоритма для предсказания рыночной стоимости CatBoost.

Введем новую целевую переменную, отвечающую за рыночную стоимость земельных участков. Переменная представляет собой разницу между стоимостью участка за квадратный метр земли в 2021 и 2015 гг. Применим алгоритм *CatBoost*, точность модели составила 95 %, как и для предсказания стоимости земельных участков по стоимости за квадратный метр.

Обсуждение

Принцип работы будущего программного продукта заключается в том, что в специальном плагине пользователь может выбрать участок на карте или задать его координаты и получить на выходе кластер, гипотетическую стоимость земельных участков и прибыль за некоторое заданное количество лет. Предельный временной промежуток предсказания рыночной стоимости составляет 5 лет. При решении задачи прогнозирования рыночной стоимости существуют проблемы, в первую очередь связанные с тем, что сделки по куплепродаже участков происходят относительно нечасто, следовательно, построение временного ряда для земельных участков затруднительно ввиду отсутствия значений стоимости во все необходимые временные точки. Восстановление временного ряда – сложная и требующая особого внимания подзадача.

Заключение

В результате исследования было установлено сильное влияние местоположения земельного участка относительно объектов социальной инфраструктуры (школ, поликлиник и т. п.), федеральных трасс (трасса М11), крупных населенных пунктов (Москва, Тверь) и водоемов (Волга).

Построена математическая модель на основе ансамблевого самооптимизирующегося метода, позволяющая определять стоимость квадратного метра земельных участков. Точность работы алгоритма составляет 95 %. Так, при опробовании алгоритма для предсказания стоимости земельного участка погрешность модели составляет 47,3 руб. при стоимости квадратного метра земельного участка 923 руб.

Исследование было проведено в рамках исполнения государственного задания Министерства науки и высшего образования РФ, госбюджетная тема № FSFE-2022-0001.

СПИСОК ЛИТЕРАТУРЫ

1. Боровков А. А. Математическая статистика : учеб. – 4-е изд., стер. – СПб. : Издательство «Лань», 2010. – 704 с.

2. Шпигельхалтер Д. Искусство статистики. – М. : Манн, Иванов и Фербер, 2020. – С. 134–155.

3. Любимцев О. В., Любимцева О. Л. Линейные регрессионные модели в эконометрике, Линейные регрессионные модели в эконометрике. – Нижний Новгород : Нижегородский государственный архитектурно-строительный университет, ЭБС АСВ, 2016. – С. 9–16.

4. Деммель Дж. Вычислительная линейная алгебра. – М. : Мир, 2000. – 430 с.

5. Vandeput N. Data Science for Supply Chain Forecasting. – 2nd ed. – Walter de Gruyter GmbH & Co KG, 2021 – 310 p.

6. Duan T., Avati A., Ding D. Yi., Thai K. K., Basu S., Ng A., Schuler A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction [Electronic resource] // Proceedings of the 37 th International Conference on Machine Learning. – Vienna, Austria, PMLR 108, 2020. – Mode of access: https://arxiv.org/pdf/1910.03225.pdf (accessed 01.03.2023).

7. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Data Mining, Inference and Prediction. – Springer. – P. 337–386, 409 [Electronic resource]. – Mode of access: https://hastie.su.domains/Papers/ESLII.pdf (дата обращения: 01.03.2023).

8. Dr Guillaume Saupin. Practical Gradient Boosting: An deep dive into Gradient Boosting in Python. – 2022. – 208 p.

9. Кашницкий Ю. С. История развития ансамблевых методов классификации в машинном обучении. – С. 4–6 [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/278019662_Istoria_razvitia_ansamblevyh_metodov_klassifikacii_v_masinnom_obucenii (дата обращения: 01.03.2023).

10. Natekin A., Knoll A. Gradient boosting machines, a tutorial [Electronic resource]. – Mode of access: https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial.

11. Peng R. D., Matsui E. The Art of Data Science. A Guide for Anyone Who Works with Data [Electronic resource]. – Mode of access: http://bedford-computing.co.uk/learning/wp-content/uploads/2016/09/artofdata-science.pdf.

12. VanderPlas J. Python Data Science Handbook [Electronic resource]. – Mode of access: https://jakevdp.github.io/PythonDataScienceHandbook/.

13. Шульгин С. Г. Отбор переменных для анализа и прогнозирования нестабильности с помощью моделей градиентного бустинга [Электронный ресурс]. – Режим доступа: https://www.socionauki.ru/upload/socionauki.ru/book/files/monitoring_sm_5/115-153.pdf.

14. Nailong Zhang. A Tour of Data Science (Chapman & Hall/CRC Data Science Series). – 2020.

15. Федотов С., Синицин Ф. Машинное обучение [Электронный ресурс]. – Режим доступа: https://academy.yandex.ru/handbook/ml/article/mashinnoye-obucheniye.

Об авторах

Алексей Викторович Курлов – заведующий научно-исследовательской лабораторией городских технологий и пространственного развития.

Получено 22.06.2023

Development of a predictive geoinformational model for market value assessment of land plots in expert system

A. V. Kurlov[1]

[1] Research Laboratory «Laboratory of Urban Technologies and Spatial Development», Moscow, Russian Federation

* e-mail: kurlov-av@yandex.ru

Abstract. When solving the task of market value evaluation, there is a problem associated with a set of data on the market value of land plots. In order to define real market value, a hypothesis was put forward that a model for predicting the land plot cadastral value would be effective for predicting an increase in the land plot market value as well. The research is aimed at determining the land plot market value, taking into account the

spatial factor. The article analyzes the data of the land plot market value based on registered purchase and sale transactions. The results of the study showed that the location of land plots in relation to social infrastructure, federal highways, large settlements and reservoirs has a strong influence on the market value. Based on this, a mathematical model has been developed that uses a self-optimizing ruble method to determine the cost per square meter of land plot. The principle of the future software product operation is that in a special plug-in, the user can select a land plot on the map or set its coordinates and get an investment attractiveness class, a hypothetical value of a land plot and profit for a given number of years.

Keywords: market value, cadastral value, analysis, forecast, regression model, expert model, CatBoost, NgBoost, XGBoost

REFERENCES

1. Borovkov, A. A. (2010). *Matematicheskaya statistika [Mathematical statistics]* (4th ed.). St. Petersburg: "Lan" Publ., 704 p. [in Russian].

2. Spiegelhalter, D. (2020). *Iskusstvo statistiki [The Art of Statistics]* (pp. 134–155). Moscow: Mann, Ivanov i Ferber Publ. [in Russian].

3. Lyubimtsev, O. V., & Lyubimtseva, O. L. (2016). Lineynye regressionnye modeli v ekonometrike, Lineynye regressionnye modeli v ekonometrike [Linear regression models in econometrics] (pp. 9–16). Nizhny Novgorod: Nizhny Novgorod State University of Architecture and Civil Engineering, EBS ASV Publ. [in Russian].

4. Demmel, J. (2000). *Vychislitel'naya lineynaya algebra [Computational linear algebra]*. Moscow: Mir Publ., 430 p. [in Russian].

5. Vandeput, N. (2021). Data Science for Supply Chain Forecasting (2nd ed.). Walter de Gruyter GmbH & Co KG, 310 p.

6. Duan, T., Avati, A., Ding, D. Yi., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020). NGBoost: Natural Gradient Boosting for Probabilistic Prediction. *Proceedings of the 37 th International Conference on Machine Learning*. Vienna, Austria, PMLR 108. Retrieved from https://arxiv.org/pdf/1910.03225.pdf (accessed 01.03.2023).

7. Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction (pp. 337–386, 409). Springer. Retrieved from https://hastie.su.domains/Papers/ESLII.pdf (accessed March 01, 2023).

8. Dr Guillaume Saupin. (2022). Practical Gradient Boosting: An deep dive into Gradient Boosting in Python (208 p.).

9. Kashnitskiy, Yu. S. History of development of ensemble classification methods in machine learning (pp. 4–6). Retrieved from https://www.researchgate.net/publication/278019662_Istoria_razvitia_ansamblevyh_metodov_klassifikacii_v_masinnom_obucenii (accessed March 01, 2023) [in Russian].

10. Natekin, A., & Knoll, A. Gradient boosting machines, a tutorial. Retrieved from https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial

11. Peng, R. D., & Matsui, E. The Art of Data Science. A Guide for Anyone Who Works with Data. Retrieved from http://bedford-computing.co.uk/learning/wp-content/uploads/2016/09/artofdatascience.pdf.

12. VanderPlas J. Python Data Science Handbook. Retrieved from https://jakevdp .github.io/PythonDataScienceHandbook/.

13. Shulgin, S. G. Selection of variables for the analysis and prediction of instability using gradient boosting models. Retrieved from https://www.socionauki.ru/upload/socionauki.ru/book/files/monitoring_sm_5/115-153.pdf.

14. Nailong Zhang. (2020). A Tour of Data Science (Chapman & Hall/CRC Data Science Series).

15. Fedotov, S., & Sinitsin, F. Machine learning. Retrieved from https://academy.yandex.ru/handbook/ml/article/mashinnoye-obucheniye [in Russian].

About the authors

Alexey V. Kurlov – Head of the Research Laboratory of Urban Technologies and Spatial Development.

Received 22.06.2023