Вестник РУДН. Серия: Инженерные исследования
RUDN Journal of Engineering Research

2024;25(2):151–161

ISSN 2312-8143 (Print); ISSN 2312-8151 (Online)
journals.rudn.ru/engineering-researches

DOI: 10.22363/2312-8143-2024-25-2-151-161
УДК 004.852
EDN: TMZJXN

Научная статья / Research article

Построение предиктивной модели для прогнозирования цен недвижимости на основе сформированной базы данных

П.А. Коняева, О.А. Салтыкова✉, С.А. Купреев

Российский университет дружбы народов, *Москва, Россия*
✉ saltykova-oa@rudn.ru

История статьи

Поступила в редакцию: 28 января 2024 г.
Доработана: 12 марта 2024 г.
Принята к публикации: 8 апреля 2024 г.

Заявление о конфликте интересов

Авторы заявляют об отсутствии конфликта интересов.

Вклад авторов

Нераздельное соавторство.

Аннотация. Представлено решение актуальной задачи прогнозирования цен на недвижимость с помощью построения предиктивной модели на основе сформированной базы данных по недвижимости в Москве, размещенной на веб-сайте «Move Недвижимость». Рассмотрены существующие методы машинного обучения для решения задачи прогнозирования и применен один из них — множественная линейная регрессия. Проведен регрессионный анализ полученных результатов решения задачи прогнозирования. В качестве управляющих параметров рассматриваются 11 независимых переменных. Исследовано влияние учитываемых при построении модели переменных на результаты решения задачи прогнозирования цен на недвижимость. Определено, какие из независимых переменных оказывают наибольшее влияние на результаты работы модели. Для улучшения качества модели была осуществлена предобработка и стандартизация признаков, а также идентификация выбросов и пропусков значений при формировании базы данных. Коэффициенты модели множественной линейной регрессии определялись с помощью метода наименьших квадратов. Для оценки качества модели проводился анализ следующих параметров модели: R-квадрат, скорректированный R-квадрат, p-значение. Результатом построения предиктивной модели является полученное уравнение регрессии. Применение полученного уравнения может быть использовано для последующего учета конкретных характеристик при решении задачи прогнозирования цен на недвижимость. Показаны преимущества использования данного метода и перспективы применения полученного результата.

Ключевые слова: машинное обучение, регрессионный анализ, множественная линейная регрессия, метод наименьших квадратов, уравнение регрессии, R-квадрат, скорректированный R-квадрат, p-значение

Для цитирования

Коняева П.А., Салтыкова О.А., Купреев С.А. Построение предиктивной модели для прогнозирования цен недвижимости на основе сформированной базы данных // Вестник Российского университета дружбы народов. Серия: Инженерные исследования. 2024. Т. 25. № 2. С. 151–161. http://doi.org/10.22363/2312-8143-2024-25-2-151-161

Building a Predictive Model for Predicting Real Estate Prices Based on the Generated Database

Polina A. Konyaeva, Olga A. Saltykova✉, Sergei A. Kupreev

RUDN University, *Moscow, Russia*
✉ saltykova-oa@rudn.ru

Article history

Received: January 28, 2024
Revised: March 12, 2024
Accepted: April 8, 2024

Conflicts of interest

The authors declare that there is no conflict of interest

Authors' contribution

Undivided co-authorship.

Abstract. The work is devoted to solving the current problem of forecasting real estate prices by building a predictive model based on the generated database of real estate in Moscow, posted on the Move Real Estate website. Existing machine learning methods for solving the forecasting problem are considered and one of them is applied — multiple linear regression. A regression analysis of the obtained results of solving the forecasting problem was carried out. Eleven independent variables are considered as control parameters. The influence of the variables taken into account when constructing the model on the results of solving the problem of forecasting real estate prices was studied. It was determined which of the independent variables have the greatest impact on the results of the model. To improve the quality of the model, preprocessing and standardization of features were carried out. Identification of outliers and omissions of values was carried out during the formation of the database. The coefficients of the multiple linear regression model were determined using the least squares method. To assess the quality of the model, the following model parameters are analyzed: *R*-squared, adjusted *R*-squared, *p*-value. The result of constructing a predictive model is the resulting regression equation. The application of the resulting equation can be used to subsequently take into account specific characteristics when solving the problem of forecasting real estate prices. The work shows the advantages of using this method and the prospects for applying the obtained result.

Keywords: machine learning, regression analysis, multiple linear regression, least squares method, regression equation, *R*-squared, adjusted *R*-squared, *p*-value

For citation

Konyaeva PA, Saltykova OA., Kupreev SA. Building a predictive model for predicting real estate prices based on the generated database. *RUDN Journal of Engineering Research.* 2024;25(2):151–161. (In Russ.) http://doi.org/10.22363/2312-8143-2024-25-2-151-161

Введение

Машинным обучением называют совокупность методов и алгоритмов, направленных на автоматическое обучение системы, основываясь на определенных данных [1]. Методы и подходы машинного обучения часто применяются к решению не только сугубо научных, но и прикладных задач, в частности для решения задач прогнозирования изменения цен на недвижимость [2–7], поведения клиентов [8–11], динамики рынка ценных бумаг [12–14] и пр. Методы, которые применяются для решения вышеупомянутых задач, относятся к методам

машинного обучения и интеллектуального анализа данных, активное развитие и применение которых мы наблюдаем в последнее время [15–16]. В литературе можно встретить работы по прогнозированию цен на московском рынке с точки зрения микро- и макроэкономических показателей [17–18]. Целью данной работы является решение актуальной задачи по построению математической модели на основе методов машинного обучения, для решения задачи прогнозирования динамики цен на недвижимость в Москве на основе данных с сайта о недвижимости. В области машинного обучения широко используется регрессионный анализ [19], одним из наиболее распространенных методов в прогнозировании цен на недвижимость является множественная линейная регрессия [20]. Для использования этого метода необходимо провести анализ нескольких переменных, которые могут повлиять на цену недвижимости. Это позволяет определить, какие из независимых переменных оказывают наибольшее влияние на цену недвижимости. Полученная в результате моделирования модель, построенная на основе реальных данных, позволит предсказать цену недвижимости на основании наиболее значимых факторов, влияющих на ее динамику.

1. Методы исследования

На основе сформированной базы данных по объявлениям недвижимости, размещенным на веб-сайте «Move Недвижимость»[1] и обработанным параметрам будет построена предиктивная модель и выведено уравнение регрессии. Для предсказания цены, а также значения и оценки связи между одной зависимой переменной (ценой) и 11 независимыми переменными (количество комнат, площадь, этаж квартиры, этаж дома, год постройки, расстояние до метро, тип транспорта до метро, тип продажи, тип квартиры, время до центра города, время до метро) будет использоваться статистический метод — множественная линейная регрессия, которую можно представить в виде следующего уравнения:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \qquad (1)$$

где y — зависимая переменная, $x_1, x_2, \ldots x_k$ — независимые переменные, β_0, $\beta_1, \beta_2, \ldots \beta_k$ — коэффициенты, а ε — случайная ошибка. Коэффициенты модели множественной линейной регрессии находятся с помощью метода наименьших квадратов[2]. Метод наименьших квадратов (МНК) — это метод оценки параметров линейной регрессии, используемый для минимизации суммы квадратов разностей между наблюдаемыми значениями и значениями, расчет которых производится по модели.

2. Реализация построения предиктивной модели

Построим первую модель МНК[3], не производя предварительную обработки данных (рис. 1).

В результате получается следующее уравнение регрессии:

$$\begin{aligned} y = {} & 148916981.57767546 + 9426604.6819443 * \text{Number rooms} + \\ & + 18.550068520072813 * \text{Square} + \\ & + 239260.53265802326 * \text{Floor sq.} + \\ & + 73329.76103038796 * \text{Floors} + (-47231.61525882067 * \text{Year}) + \\ & + 24804.15186899586 * \text{Distance metro km} + \\ & + (-1487711.8963242061 * \text{Type transport to metro}) + \\ & + (-13554120.797380297 * \text{Sale type}) + \\ & + (-41923145.71242697 * \text{Flat type}) + \\ & + (-406837.02772176114 * \text{Time to the center (from metro)}) + \\ & + 61407.48561820931 * \text{Time to metro in minutes}. \end{aligned} \qquad (2)$$

[1] Move Недвижимость // Move.ru — портал о недвижимости Москвы. URL: https://move.ru/ (дата обращения: 28.02.2023)

[2] Профессиональное сообщество NTA Метод наименьших квадратов: формулы, код и применение // Хабр. 2022. URL: https://habr.com/ru/articles/672540/ (дата обращения: 09.04.2023).

[3] Как читать и интерпретировать таблицу регрессии // Кодкамп. 2022. URL: https://www.codecamp.ru/blog/read-interpret-regression-table/ (дата обращения: 10.04.2023).

Далее будут просмотрены параметры полученной модели, но перед этим определим, что такое *R*-квадрат (*R*-squared), скорректированный *R*-квадрат (Adj. *R*-squared), *p*-значение (*P* для коэффициентов и Prob (*F*-statistic) для всей модели)[4].

R-квадрат является статистической мерой, которая оценивает, насколько хорошо выбранная модель подходит для наблюдаемых данных и определяется следующей математической формулой:

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{\frac{SS_{res}}{n}}{\frac{SS_{tot}}{n}} = 1 - \frac{SS_{res}}{SS_{tot}}, \qquad (3)$$

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.096
Model:	OLS	Adj. R-squared:	0.096
Method:	Least Squares	F-statistic:	391.0
Date:	Sun, 30 Apr 2023	Prob (F-statistic):	0.253
Time:	13:15:23	Log-Likelihood:	-7.6250e+05
No. Observations:	40597	AIC:	1.525e+06
Df Residuals:	40585	BIC:	1.525e+06
Df Model:	11		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.489e+08	2.15e+07	6.924	0.000	1.07e+08	1.91e+08
x1	9.427e+06	2.21e+05	42.586	0.000	8.99e+06	9.86e+06
x2	18.5501	18.074	1.026	0.305	-16.874	53.975
x3	2.393e+05	3.36e+04	7.120	0.021	1.73e+05	3.05e+05
x4	7.333e+04	2.45e+04	2.988	0.003	2.52e+04	1.21e+05
x5	-4.723e+04	1.06e+04	-4.472	0.000	-6.79e+04	-2.65e+04
x6	2.48e+04	8.06e+04	0.308	0.758	-1.33e+05	1.83e+05
x7	-1.488e+06	5.29e+05	-2.810	0.005	-2.53e+06	-4.5e+05
x8	-1.355e+07	5.22e+05	-25.947	0.214	-1.46e+07	-1.25e+07
x9	-4.192e+07	2.89e+06	-14.481	0.030	-4.76e+07	-3.62e+07
x10	-4.068e+05	1.56e+04	-26.058	0.000	-4.37e+05	-3.76e+05
x11	6.141e+04	1.93e+04	3.185	0.401	2.36e+04	9.92e+04

Рис. 1. Модель 1
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 1. Model 1
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

где $SS_{res} = \Sigma_{i=1}^{n} e_i^2 = \Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$ — сумма квадратов остатков регрессии; $y_i, \hat{y}_i$ — фактические и расчетные значения объясняемой переменной; $SS_{tot} = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2 = n\hat{\sigma}_y^2$ — общая сумма квадратов; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. В случае линейной регрессии с константой:

$$SS_{tot} = SS_{reg} + SS_{res},$$

где $SS_{reg} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2$ — объясненная сумма квадратов. Отсюда возможно получить упрощенное определение *R*-квадрата как доли объясненной суммы квадратов в общей[6]:

$$R^2 = \frac{SS_{reg}}{SS_{tot}}. \qquad (4)$$

R-квадрат может принимать значения от 0 до 1. Значение 0 означает, что выбранная модель не объясняет никакой изменчивости

[4] *R* — значит регрессия // Хабр. 2018. URL: https://habr.com/ru/articles/350668/ (дата обращения: 10.04.2023).
[6] Коэффициент детерминации. // Википедия. 2022. URL: https://ru.wikipedia.org/wiki/Коэффициент_детерминации (дата обращения: 10.04.2023).

наблюдаемых данных, а значение 1 указывает на идеальное соответствие выбранной модели и наблюдаемых данных.

Скорректированный *R*-квадрат, или скорректированный коэффициент детерминации, является статистической мерой, которая учитывает количество регрессоров (независимых переменных) в модели, и представляет собой измененный коэффициент детерминации *R*-квадрат, исправленный с учетом числа независимых переменных в модели:

$$R^2_{\text{скорр}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k-1}, \quad (5)$$

где R^2 — коэффициент детерминации, n — количество наблюдений (то есть размер выборки), k — количество независимых переменных в модели.

С помощью скорректированного *R*-квадрата можно оценить, насколько хорошо модель соответствует данным, учитывая количество регрессоров. Чем ближе значение скорректированного *R*-квадрата к 1, тем лучше соответствие модели данным.

P-значение — это вероятность получения результата, не менее экстремального, чем наблюдаемый результат, при условии, что нулевая гипотеза верна. Если *P*-значение меньше предварительно определенного уровня значимости (менее 0,05), то нулевая гипотеза отвергается, что означает, что существует статистически значимая разница между группами или переменными. В противном случае (*P* более 0,05) нулевая гипотеза не отвергается, что означает, что недостаточно данных, чтобы вынести окончательный вывод.

Проверим значимость коэффициентов регрессии β_k. Для этого выведем гипотезы[7] для β_k для $k = \overline{0;11}$:

H0: $\beta_k = 0$ — коэффициент незначим;

H1: $\beta_k \neq 0$ — коэффициент значим;

$\alpha = 0{,}05$,

P-значение у коэффициентов β_k = (0,000; 0,000; 0,021; 0,003; 0,000; 0,005; 0,030; 0,000) < $< \alpha = 0{,}05$, поэтому принимаем гипотезу H1 (коэффициент β_k значим с вероятностью 95 % для k = 0, 1, 3, 4, 5, 7, 9, 10).

P-значение у коэффициентов β_k = (0,305; 0,758; 0,214; 0,401) > α = 0,05, поэтому принимаем гипотезу H0 (коэффициент β_k незначим с вероятностью 95 % для k = 2, 6, 8, 11).

Значения параметров, полученные для всей модели:

R-squared = 0,096; Adj. *R*-squared = 0,096; Prob (*F*-statistic) = 0,253.

Вывод: уравнение незначимо при уровне значимости $\alpha = 0{,}05$ (0,253 > 0,05), при этом 9,6 % (и 9,6 % по скорректированному *R*-квадрату) — вариация целевой переменной y = price объясняется вариацией факторов.

На следующем этапе после построения первой модели и получения результатов параметров необходимо сделать обработку данных. При проверке данных возможно столкнуться с выбросами[8] (рис. 2) — аномальными значениями, которые необходимо убрать. После предобработки данных (рис. 3) будут построены гистограммы без аномальных значений.

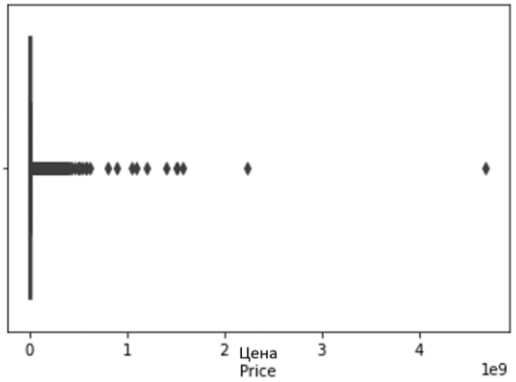

Рис. 2. Выброс в данных с параметром — цена
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 2. Outlier in data with the price parameter
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

[7] Мир статистических гипотез // Хабр. 2021. URL: https://habr.com/ru/articles/558836/ (дата обращения: 11.04.2023).

[8] Выбросы в данных // Машинное обучение. 2023. URL: https://www.dmitrymakarov.ru/data-analysis/outliers-09/ (дата обращения: 11.04.2023).

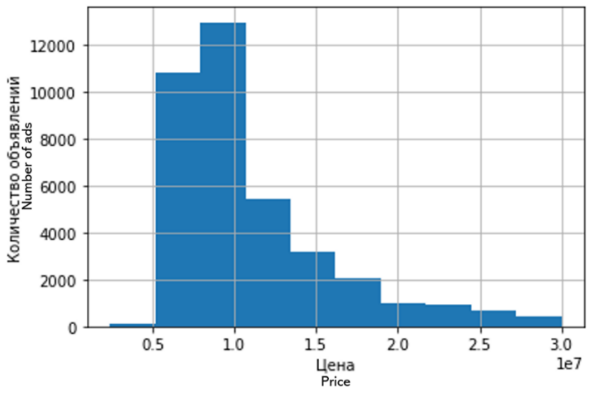

Рис. 3. Гистограмма по цене после удаления выбросов
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 3. Histogram by price after removal of outliers
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

Работу с выбросами необходимо проделать со всеми параметрами из базы данных. После обработки построим вторую модель МНК (рис. 4).

В результате получается следующее уравнение регрессии:

$$
\begin{aligned}
y = {} & -13143953.557547348 + \\
& + 1746037.573731812*\text{Number rooms} + \\
& 41982.3071491594*\text{Square} + 46420.9958310878*\text{Floor sq.} + \\
& + 79133.40581343621*\text{Floors} + 12236.46452122052*\text{Year} + \\
& + (-153413.0293165489*\text{Distance metro km}) + \\
& + (-57272.23972169444*\text{Type transport to metro}) + \\
& + (-5411948.059501247*\text{Sale type}) + \\
& + (-2621892.3884117184*\text{Flat type}) + \\
& + (-60224.83508604619*\text{Time to the center (from metro)}) + \\
& + (16731.62320475588*\text{Time to metro in minutes}).
\end{aligned}
\tag{6}
$$

Аналогично проверим значимость коэффициентов регрессии βk. *P*-значение у коэффициентов βk = (0,000; 0,000; 0,007; 0,000; 0,000; 0,000; 0,000; 0,014; 0,000; 0,000; 0,000; 0,005) < $< \alpha = 0{,}05$, поэтому принимаем гипотезу H1 (коэффициент βk значим с вероятностью 95 % для $k = \overline{0; 11}$).

Значения параметров, полученные для всей модели: *R*-squared = 0,864; Adj. *R*-squared = 0,857; Prob (*F*-statistic) = 0,000.

Вывод: уравнение значимо при α = 0,05.

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.864
Model:	OLS	Adj. R-squared:	0.857
Method:	Least Squares	F-statistic:	5496.
Date:	Sun, 30 Apr 2023	Prob (F-statistic):	0.000
Time:	13:16:04	Log-Likelihood:	-4.9302e+05
No. Observations:	38568	AIC:	9.861e+05
Df Residuals:	38568	BIC:	9.862e+05
Df Model:	11		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.314e+07	1.58e+06	-8.321	0.000	-1.62e+07	-1e+07
x1	1.746e+06	3.05e+04	57.188	0.000	1.69e+06	1.81e+06
x2	4.198e+04	1349.681	31.105	0.007	3.93e+04	4.46e+04
x3	4.642e+04	3190.887	14.548	0.000	4.02e+04	5.27e+04
x4	7.913e+04	2324.085	34.049	0.000	7.46e+04	8.37e+04
x5	1.224e+04	766.854	15.957	0.000	1.07e+04	1.37e+04
x6	-1.534e+05	6963.841	-22.030	0.000	-1.67e+05	-1.4e+05
x7	-5.727e+04	4.51e+04	-1.269	0.014	-1.46e+05	3.12e+04
x8	-5.412e+06	4.33e+04	-124.965	0.000	-5.5e+06	-5.33e+06
x9	-2.622e+06	2.97e+05	-8.840	0.000	-3.2e+06	-2.04e+06
x10	-6.022e+04	1289.406	-46.707	0.000	-6.28e+04	-5.77e+04
x11	1.673e+04	2779.242	6.020	0.005	1.13e+04	2.22e+04

Рис. 4. Модель 2
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 4. Model 2
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

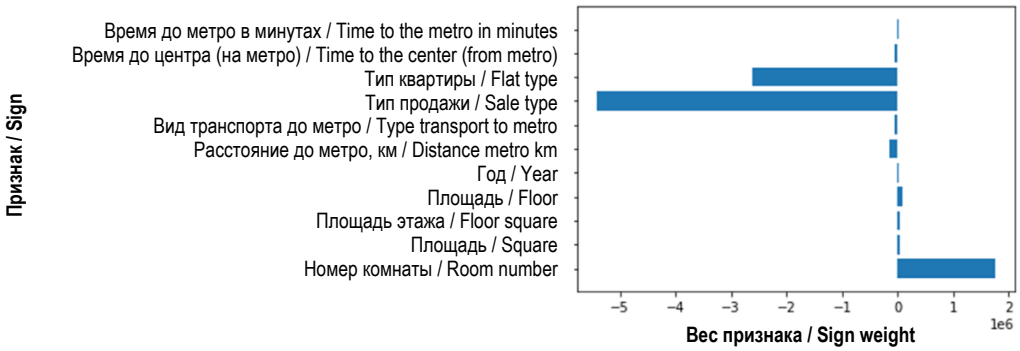

Рис. 5. Визуализация значимости признаков
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 5. Visualization of the significance of signs
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

Произведем визуализацию того, какой вклад вносит каждый из коэффициентов (рис. 5). Очевидно, что признаки имеют совершенно разное влияние на итоговую цену. При просмотре базы с параметрами, содержащей статистические данные, убедимся, что значения по каждому признаку имеют разный масштаб разброса значений. Для решения этой проблемы можно применить стандартизацию признаков (это процесс преобразования значений признаков в наборы с нулевым средним и единичным стандартным отклонением) [21]. Для этого необходимо от значений каждого признака отнять среднее значение этого признака, а затем поделить на среднее квадратическое отклонение этого признака:

$$x_{\text{scaled}} = \frac{x - x_{\text{mean}}}{\sigma_x}. \qquad (7)$$

После стандартизации каждый признак имеет среднее значение 0 и среднее квадратическое отклонение 1. Построим третью модель на стандартизированных данных (рис. 6).

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.917
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	5496.
Date:	Mon, 01 May 2023	Prob (F-statistic):	0.000
Time:	13:03:04	Log-Likelihood:	-4.9302e+05
No. Observations:	38568	AIC:	9.861e+05
Df Residuals:	38568	BIC:	9.862e+05
Df Model:	11		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.068e+07	1.39e+04	768.249	0.000	1.06e+07	1.07e+07
x1	1.277e+06	2.23e+04	57.188	0.000	1.23e+06	1.32e+06
x2	6.884e+05	2.21e+04	31.105	0.000	6.45e+05	7.32e+05
x3	2.622e+05	1.8e+04	14.548	0.000	2.27e+05	2.98e+05
x4	6.528e+05	1.92e+04	34.049	0.000	6.15e+05	6.9e+05
x5	2.642e+05	1.66e+04	15.957	0.000	2.32e+05	2.97e+05
x6	-4.928e+05	2.24e+04	-22.030	0.000	-5.37e+05	-4.49e+05
x7	-2.771e+04	2.18e+04	-1.269	0.005	-7.05e+04	1.51e+04
x8	-2.527e+06	2.02e+04	-124.965	0.000	-2.57e+06	-2.49e+06
x9	-1.235e+05	1.4e+04	-8.840	0.000	-1.51e+05	-9.61e+04
x10	-7.284e+05	1.56e+04	-46.707	0.000	-7.59e+05	-6.98e+05
x11	9.06e+04	1.5e+04	6.020	0.000	6.11e+04	1.2e+05

Рис. 6. Модель 3
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой
Figure 6. Model 3
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

В результате получается следующее уравнение регрессии:

$$
\begin{aligned}
y = {} & 10675976.805853501 + \\
& + 1277111.325193666*\text{Number rooms} + \\
& + 688409.5206161594*\text{Square} + \\
& 262237.3861052125*\text{Floor sq.} + \\
& + 652771.7247589253*\text{Floors} + 264218.82554299495*\text{Year} + \\
& (-492783.3576360492*\text{Distance metro km}) + \\
& + (-27706.89877696723*\text{Type transport to metro}) + \\
& + (-2526546.283715316*\text{Sale type}) + \\
& + (-123500.06207640213*\text{Flat type}) + \\
& + (-728416.9402160242*\text{Time to the center (from metro)}) + \\
& + 90596.00852600564*\text{Time to metro in minutes}
\end{aligned}
\qquad (8)
$$

Аналогично проверим значимость коэффициентов регрессии β_k. *P*-значение у коэффициентов

β_k = (0,000; 0,000; 0,000; 0,000; 0,000; 0,000; 0,000; 0,005; 0,000; 0,000; 0,000; 0,000) < α = 0,05.

Так же как и в предыдущей модели, принимаем гипотезу H1 (коэффициент β_k значим с вероятностью 95 % для $k = \overline{0; 11}$).

Значения параметров улучшились:

R-squared = 0,917; Adj. *R*-squared = 0,898; Prob (*F*-statistic) = 0,000.

Вывод: уравнение значимо при уровне значимости α = 0,05 (0,000 < 0,05).

Данная модель улучшилась и будет предлагаться для использования предсказания цен на недвижимость, поставляя необходимые параметры.

Для новой окончательной модели также произведем визуализацию значимости признаков (рис. 7).

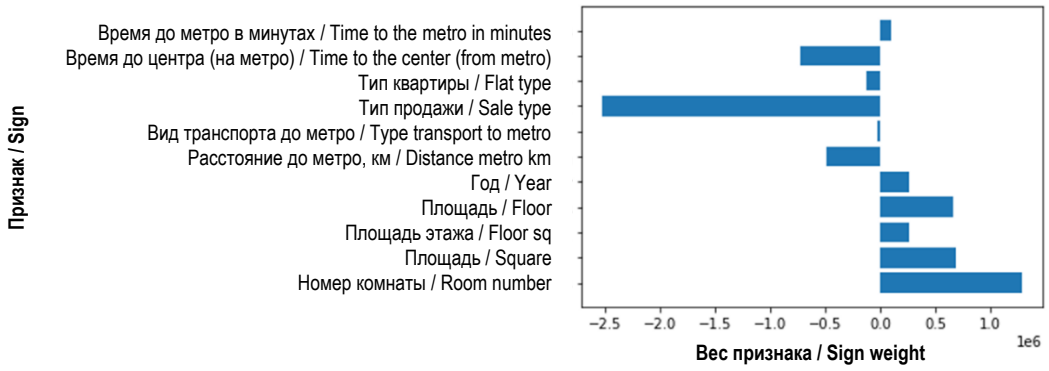

Рис. 7. Визуализация значимости признаков для новой модели
И с т о ч н и к : выполнено П.А. Коняевой, О.А. Салтыковой

Figure 7. Visualization of the significance of signs for the new model
S o u r c e : compiled by P.A. Konyaeva, O.A. Saltykova

После стандартизации данных веса признаков имеют совершенно иные значения относительно друг друга. Важно отметить, что стандартизация важна не только для отбора признаков, но также является важным этапом предобработки данных, без которого многие алгоритмы будут работать некорректно.

В результате после удаления выбросов и стандартизации данных получена улучшенная предиктивная модель. Также получено уравнение регрессии, предназначенное для прогнозирования цен недвижимости с указанием конкретных параметров, на основе сформированной базы данных.

Заключение

В работе представлено построение предиктивной модели с помощью метода наименьших квадратов на основе сформированной базы данных. Для улучшения модели была осуществлена предобработка и стандартизация

признаков, после чего построено уравнение регрессии для возможности прогнозирования и анализа развития цен на рынке и оценки недвижимости при внесении необходимых параметров.

Собранная в данной работе информация может помочь отследить и сравнить цены на квартиры, дать общую картину рынка недвижимости за определенный период и определить наилучший вариант жилья для предоставления клиенту. Используя полученный результат в виде построения модели, возможно рассчитать состояние рынка недвижимости и построить прогнозы о его дальнейшем развитии, что, в свою очередь, предоставляет возможность для наиболее выгодных инвестиционных вложений. Вместе с тем полученная информация будет полезна для агентств недвижимости, что принесет дополнительную прибыль.

Список литературы

1. *Алексеев Г.* Введение в машинное обучение // Хабр. 2019. URL: https://habr.com/ru/articles/448892/ (дата обращения: 27.03.2023).

2. *Лейфер Л.А., Чёрная Е.В.* Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир // Имущественные отношения в Российской Федерации. 2020. №. 3 (222). С. 32–42. EDN: BQRFXJ

3. *Kok N., Koponen E.-L., Martinez-Barbosa C.A.* (2017). Big Data in Real Estate From Manual Appraisal to Automated Valuation» // The Journal of Portfolio Management. 2017. Vol. 43. No. 6. P. 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

4. *Ясницкий В.Л.* Нейросетевое моделирование в задаче массовой оценки жилой недвижимости города Перми // Фундаментальные исследования. 2015. № 10–3. С. 650–653. EDN: UNXWSX

5. *Сурков Ф.А., Петкова Н.В., Суховский С.Ф.* Нейросетевые методы анализа данных в оценке недвижимости» // Известия вузов. Северо-Кавказский регион. Технические науки. 2016. № 3. С. 38–45. https://doi.org/10.17213/0321-2653-2016-3-38-45

6. *Арефьева Е.А., Костяев Д.С.* Использование нейронных сетей для оценки рыночной стоимости недвижимости // Известия Тульского государственного университета. Технические науки. 2017. № 10. С. 177–184. EDN: ZVLGJH

7. *Выходцев Н.А.* Использование искусственного интеллекта для оценки стоимости недвижимого имущества // Доклады Томского государственного университета систем управления и радиоэлектроники. 2021. Т. 24. № 1. С. 68–72. https://doi.org/10.21293/1818-0442-2021-24-1-68-72

8. *Арзамасцев С.А., Бгатов М.В., Картышева Е.Н., Деркунский В.А., Семенчиков Д.Н.* Предсказание оттока абонентов: сравнение методов машинного обучения // Компьютерные инструменты в образовании. 2018. № 5. С. 5–23. https://doi.org/10.32603/2071-2340-2018-5-5-23

9. *Радчук М. А., Копытина Е.А.* Разработка программного средства для предсказания оттока клиентов с помощью методов машинного обучения // Сборник студенческих научных работ факультета компьютерных наук ВГУ. 2019. С. 190–196. EDN: PSWAXM

10. *Lalwani P., Mishra M.K., Chadha J.S., Sethi P.* Customer churn prediction system: a machine learning approach // Computing. 2022. Vol. 104. No. 2. P. 271–294. https://doi.org/10.1007/s00607-021-00908-y

11. *Khodabandehlou S., Zivari Rahman M.* Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior // Journal of Systems and Information Technology. 2017. Vol. 19. Iss. 1/2. P. 65–93. https://doi.org/10.1108/JSIT-10-2016-0061

12. *Андрианова Е.Г., Новикова О.А.* Роль методов интеллектуального анализа текста в автоматизации прогнозирования рынка ценных бумаг // Cloud of science. 2018. Т. 5. № 1. С. 196–211. EDN: YUTIIN

13. *Коваленко И.А.* Использование искусственного интеллекта на биржевом и внебиржевом рынке ценных бумаг // Вестник науки. 2023. Т. 3. № 6 (63). С. 75–80. URL: https://www.xn----8sbempclcwd3bmt.xn--p1ai/article/8956 (дата обращения: 28.02.2023).

14. *Henrique B.M., Sobreiro V.A., Kimura H.* Literature review: Machine learning techniques applied to financial market prediction // Expert Systems with Applications. 2019. Vol. 124. P. 226–251. https://doi.org/10.1016/j.eswa.2019.01.012

15. *Kumbure M.M., Lohrmann C., Luukka P., Porras J.* Machine learning techniques and data for stock market forecasting: A literature review // Expert Systems with Applications. 2022. Vol. 197. https://doi.org/10.1016/j.eswa.2022.116659

16. *Mahesh B.* Machine learning algorithms-a review // International Journal of Science and Research. 2020. Vol. 9. Iss. 1. P. 381–386. https://doi.org/10.21275/ART20203995

17. *Сальников В.А., Михеева О.М.* Модели прогнозирования цен на московском рынке жилой недвижимости // Проблемы прогнозирования. 2018. № 1 (166). С. 129–139. EDN: YLXJZZ

18. *Стерник Г.М., Печенкина А.В.* Прогноз цен предложения квартир на московском рынке жилья (макроэкономический подход) // Имущественные отношения в Российской Федерации. 2007. № 10. С. 11–18. EDN: JXADIB

19. *Назаров А*. Регрессионный анализ в Data Science. Простая линейная регрессия. Библиотека statsmodels // Хабр. 2022. URL: https://habr.com/ru/articles/690414/ (дата обращения: 30.03.2023).

20. *Дронов В.* Линейная регрессия с помощью Scikit-Learn в Python // Обучение Python. 2021. URL: https://tonais.ru/library/lineynaya-regressiya-s-pomoschyu-scikit-learn-v-python (дата обращения: 05.04.2023).

21. *Айлин А.* Нормализация против стандартизации в линейной регрессии // Машинное обучение. 2023. URL: https://www.baeldung.com/cs/normalization-vs-standardization (дата обращения: 15.04.2023).

References

1. Alekseev G. Introduction to machine learning. *Habr.* 2019. (In Russ.) Available from: https://habr.com/ru/articles/448892 / (accessed: 03.27.2023).

2. Leifer LA, Chernaya EV. Machine learning techniques for real estate mass valuation. Analysis of accuracy for various methods on the example of the appraisal of apartments. *Property relations in the Russian Federation*. 2020;3:32–42. (In Russ.) EDN: BQRFXJ

3. Kok N, Koponen E-L, Martinez-Barbosa CA. Big Data in Real Estate From Manual Appraisal to Automated Valuation». *The Journal of Portfolio Management*. 2017; 43(6):202–211. https://doi.org/10.3905/jpm.2017.43.6.202

4. Yasnitskiy VL. Using a neural network to solve the problem of mass real estate appraisal of city Perm. *Fundamental Research*. 2015;10–3:650–653. (In Russ.) EDN: UNXWSX

5. Surkov FA, Petkova NV, Sukhovskiy SF. Neural network data analysis methods in real estate valuation. *News of universities. North Caucasus region. Technical science*. 2016;3:38–45. (In Russ.) https://doi.org/10.17213/0321-2653-2016-3-38-45

6. Arefieva E.A, Kostyaev D S. Using neural NETWORKS for evaluation of market cost of real estate. *News of the Tula State University. Technical science*. 2017; 10:177–184. (In Russ.) EDN: ZVLGJH

7. Vykhodtsev NA. Artificial intelligence in price estimation of real estate. *Proceedings of the TUSUR University*. 2021;24(1):68–72. (In Russ.) https://doi.org/10.21293/1818-0442-2021-24-1-68-72

8. Arzamastsev SA, Bgatov MV, Kartysheva EN, Derkunsky VA, Semenchikov DN. Predicting subscriber churn: comparison of machine learning methods. *Computer tools in education.* 2018;5:5–23. (In Russ.) https://doi.org/10.32603/2071-2340-2018-5-5-23

9. Radchuk MA, Kopytina EA. Development of a software tool for predicting customer churn using machine learning methods. *Collection of student scientific works of the Faculty of Computer Science of VSU.* 2019. p. 190–196. (In Russ.) EDN: PSWAXM

10. Lalwani P, Mishra MK, Chadha JS, Sethi P. Customer churn prediction system: a machine learning approach. *Computing*. 2022;104(2):271–294. https://doi.org/10.1007/s00607-021-00908-y

11. Khodabandehlou S, Zivari Rahman M. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology.* 2017;19(1/2):65–93. https://doi.org/10.1108/JSIT-10-2016-0061

12. Andrianova EG, Novikova OA. The role of text mining methods in automating stock market forecasting. *Cloud of science.* 2018;5(1):196–211. (In Russ.) EDN: YUTIIN

13. Kovalenko IA. Use of artificial intelligence in the exchange and over-the-counter securities markets. *Bulletin of Science.* 2023;3(6):75–80. (In Russ.) Available from: https://www.xn----8sbempclcwd3bmt.xn--p1ai/article/8956 (accessed: 30.03.2023).

14. Henrique BM, Sobreiro VA, Kimura H. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*. 2019; 124:226–251. https://doi.org/10.1016/j.eswa.2019.01.012

15. Kumbure MM, Lohrmann C, Luukka P, Porras J. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*. 2022;197:116659.

16. Mahesh B. Machine learning algorithms-a review. *International Journal of Science and Research.* 2020; 9(1):381–386. https://doi.org/10.21275/ART20203995

17. Salnikov VA, Mikheeva OM. Models for forecasting prices on the Moscow residential real estate market. *Problems of forecasting.* 2018;1(166):129–139. (In Russ.) EDN: YLXJZZ

18. Sternik GM, Pechenkina AV. Forecast of supply prices for apartments on the Moscow housing market (macroeconomic approach). *Property relations in the Russian Federation.* 2007;10:11–18. (In Russ.) EDN: JXADIB

19. Nazarov A. Regression analysis in DataScience. Simple linear regression. statsmodels library. *Habr.* 2022. (In Russ.) Available from: https://habr.com/ru/articles/690414 / (accessed: 30.03.2023).

20. Dronov V. Linear regression using Scikit-Learn in Python. *Learning Python.* 2021. (In Russ.) Available from: https://tonais.ru/library/lineynaya-regressiya-s-pomoschyu-scikit-learn-v-python (accessed: 05.04.2023).

21. Aylin A. Normalization vs. standardization in linear regression. *Machine learning.* 2023. Available from: https://www.baeldung.com/cs/normalization-vs-standardization (accessed: 15.04.2023).

Сведения об авторах

Коняева Полина Александровна, магистрант инженерной академии, Российский университет дружбы народов, Москва, Россия; E-mail: 1032212116@pfur.ru

Салтыкова Ольга Александровна, кандидат физико-математических наук, доцент департамента механики и процессов управления, инженерная академия, Российский университет дружбы народов, Москва, Россия; eLIBRARY SPIN-code: 3969-6701, ORCID: 0000-0002-3880-6662; E-mail: saltykova-oa@rudn.ru

Купреев Сергей Алексеевич, доктор технических наук, профессор департамента механики и процессов управления инженерной академии, Российский университет дружбы народов, Москва, Россия; eLIBRARY SPIN-код: 2287-2902, ORCID: 0000-0002-8657-2282; E-mail: kupreev-sa@rudn.ru

About the authors

Polina A. Konyaeva, Master's student, Academy of Engineering, RUDN University, Moscow, Russia; E-mail: 1032212116@pfur.ru

Olga A. Saltykova, Doctor of Sciences (Techn.), Associate Professor of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, Moscow, Russia; eLIBRARY SPIN-code: 3969-6701, ORCID: 0000-0002-3880-6662; E-mail: saltykova-oa@rudn.ru

Sergei A. Kupreev, Doctor of Sciences (Techn.), Professor of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, Moscow, Russia; eLIBRARY SPIN-code: 2287-2902, ORCID: 0000-0002-8657-2282; E-mail: kupreev-sa@rudn.ru