УДК 519.237
DOI 10.51691/2541-8327_2023_6_17

# *СОВМЕСТНОЕ ПРИМЕНЕНИЕ КЛАСТЕРНОГО И РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ ПРЕДСКАЗАНИЯ СТОИМОСТИ НЕДВИЖИМОСТИ*

***Базюта А.С.***
*студент,*
*Российский университет транспорта (МИИТ),*
*Москва, Россия*
***Иванова А.П.***
*к.ф.-м.н., доцент*
*Российский университет транспорта (МИИТ),*
*Москва, Россия*

**Аннотация:**
В статье исследовались возможности улучшения точности предсказания стоимости жилой недвижимости в г. Москве с помощью совместного применения множественного регрессионного анализа и кластерного анализа. Была написана программа на языке Python. Использованы открытые данные с сайта «Авито». В результате получено, что предварительное разбиение выборки на кластеры улучшает предсказательную точность регрессионной модели.
**Ключевые слова:** регрессионная модель, рейтинг объекта, кластерный анализ, рынок недвижимости, прогнозирование стоимости.

# *JOINT APPLICATION OF CLUSTER AND REGRESSION ANALYSIS TO IMPROVE THE ACCURACY OF PREDICTING THE VALUE OF REAL ESTATE*

***Bazyuta A.S.***
*student,*

*Russian University of Transport (MIIT),*
*Moscow, Russia*
***Ivanova A.P.***
*Ph.D., Associate Professor*
*Russian University of Transport (MIIT),*
*Moscow, Russia*

**Abstract:** The article investigated the possibilities of improving the accuracy of forecasting the cost of residential real estate in Moscow by using the combined application of multiple regression analysis and cluster analysis. A program was written in Python. Open data from the Avito website was used. As a result, it is found that preliminary partitioning of the sample into clusters improves the predictive accuracy of the regression model.
**Keywords:** regression model, the rating of the object, cluster analysis, real estate market, cost forecasting.

В современном мире рынок недвижимости очень динамичен. Он вызывает интерес как у юридических, так и физических лиц. Российский рынок недвижимости, как сектор национальной экономики, имеет колоссальное значение, так как является весомой составляющей ВВП и приносит в бюджет высокие доходы и налоговые поступления. На стоимость квартиры влияет множество факторов, которые имеют разную степень значимости. Чтобы принять взвешенное решение при совершении сделки с недвижимостью, нужно проанализировать большое количество исходных данных.

Во многих работах (см., например, [3,8]) используют регрессионный анализ. Мы будем применять модель множественной регрессии [2,10], имеющую вид

$$Y = X\beta + \varepsilon, \quad (1)$$

где $Y = (y_1, y_n, \dots, y_n)^T$ – вектор значений зависимой переменной, $X = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots & x_{1p} \\ 1 & x_{21} & x_{22} \cdots & x_{2p} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ 1 & x_{n1} & x_{n2} \cdots & x_{np} \end{pmatrix}$ – матрица значений объясняющих переменных размера $n \times (p+1)$, $\beta = \left(\beta_0, \beta_1, \dots, \beta_p\right)^T$ – вектор коэффициентов размера $(p+1)$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ – вектор возмущений (случайных ошибок) размера $n$.

Оценкой модели (1) по выборке будет уравнение

$$\hat{Y} = Xb + e, \qquad (2)$$

где $b = \left(b_0, b_1, \dots, b_p\right)^T$ – параметры, определяющиеся методом наименьших квадратов; $e = (e_1, e_2, \dots, e_n)^T$.

Для исследования датасет, собранный с сайта интернет-сервиса «Авито» [1,9], был разделён на тренировочную и тестовую выборки в отношении 4:1. Их размеры (433×20) и (109×20), соответственно.

Переменной *n* будет обозначаться количество строк в обучающей выборке.

В качестве зависимой переменной *Y* была выбрана цена, а в качестве объясняющих переменных *x1, x2, x3, x4, x5, x6, x7, x8, x9* – общая площадь, жилая площадь, площадь кухни, количество комнат, расстояние до центра города, время пешего пути до метро, год постройки дома, этаж, общее количество этажей в доме.

Полученное уравнение множественной регрессии имеет вид:

$$\hat{y} = -61391864.87 + 256739.53\,x1 + 119612.04\,x2 + 24040.72\,x3 +$$
$$+ (-1227278.95)x4 + (-654915.73)x5 + (-83788.86)x6 + 34686.94\,x7 +$$
$$+ 40123.22\,x8 + 175714.76\,x9.$$

Получается, что каждый квадратный метр площади квартиры увеличивает её стоимость на 256 739.53 рубля, а каждый километр расстояния от центра – уменьшает на 654 915.73 рубля.

Значимыми коэффициентами в полученном уравнении являются все коэффициенты, кроме коэффициентов, соответствующих площади кухни и

этажу. Уравнение является значимым по критерию Фишера-Снедекора [2,10]. Вариация исследуемой результирующей переменной – стоимости квартиры – на 77% обусловлена изменчивостью объясняющих переменных.

В работе также была построена модель без незначимых коэффициентов, а также была исследована регрессионная модель без мультиколлинеарности.

На основании той же выборки была построена модель множественной регрессии с переменной структурой [2,10], которая имеет вид:

$$Y = X\beta + Z\alpha + \varepsilon, \quad (3)$$

где $Z = \begin{pmatrix} 1 & z_{11} & z_{12} \cdots & z_{1m} \\ 1 & z_{21} & z_{22} \cdots & z_{2m} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ 1 & z_{n1} & z_{n2} \cdots & z_{nm} \end{pmatrix}$ – матрица значений фиктивных (структурных) переменных размера $n \times m$, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m)^T$ – вектор коэффициентов размера $m$.

В качестве бинарных переменных были выбраны *z1*, *z2*, *z3*, *z4*, *z5* – этаж, тип дома, наличие ремонта, продавец, готовность дома.

Полученное уравнения множественной регрессии с переменной структурой имеет вид:

$$\hat{y} = -8615086.98 + 226931.15\,x1 + 152789.41\,x2 + 48760.08\,x3 +$$
$$+ (-1250812.72)x4 + (-538763.8)x5 + (-83055.39)x6 + 6770.44\,x7 +$$
$$+ 60403.26\,x8 + 163973.65\,x9 + 431224.55\,z1 + 2113703.18\,z2 +$$
$$+ 2124561.64\,z3 + 793148.5\,z4 + (-2346250.65)\,z5.$$

Из полученного уравнения видно, что наличие ремонта увеличивает стоимость квартиры на 2 124 561.64 рубля, наличие посредника при сделке – на 793 148.5 рубля.

Из пяти новых коэффициентов значимыми оказались только два, соответствующие фиктивным переменным этажа и продавца. Уравнение множественной регрессии значимо по критерию Фишера-Снедекора [2,10]. Коэффициент детерминации увеличился, что говорит о влиянии добавленных в модель структурных переменных на стоимость квартиры.

Рынок недвижимости имеет неоднородную структуру. Поэтому для построения более эффективной регрессионной модели возникает необходимость выборку разбить на кластеры.

Пусть задана матрица $A = (a_{ij}),\ i = \overline{1,m}, j = \overline{1,n}$. Каждый из $m$ объектов множества $X = \{1,2,\dots,m\}$ имеет $n$ характеристик – числовых неотрицательных значений. Матрица $A$ не имеет строк и столбцов, все элементы которых равны нулю.

Каждой характеристике соответствует свой весовой коэффициент $v_j \geq 0$. Весовые коэффициенты нормируются:

$$w_j = \frac{v_j}{\sum_{j=1}^{n} v_j}, \tag{4}$$

при этом сумма всех нормированных весовых коэффициентов равна 1.

Учитывая $w_j$, вычисляются новые значения показателей:

$$b_{ij} = a_{ij} w_j, \tag{5}$$

которые тоже нормируются:

$$x_{ij} = \frac{b_{ij}}{b_j}, \tag{6}$$

где $b_j = \max\limits_{1 \leq i \leq m} b_{ij}$. Полученные значения $x_{ij}$ не имеют размерности, $0 \leq x_{ij} \leq 1$.

Пусть число $p$ непересекающихся подмножеств, на которые надо разделить множество $X$, задано. Тогда задачу можно решить, используя алгоритм «идеальной точки». Опишем его [6,7].

В каждом столбце матрицы $A = (a_{ij})$ выберем максимальный элемент: $x_j = \max\limits_{1 \leq i \leq m} x_{ij}$. Вектор $x = (x_1, x_2, \dots, x_n)$ будет соответствовать «идеальному» объекту. Для каждого объекта из множества $X$ найдём расстояние до «идеального» по формуле:

$$L_i = \max_{1 \leq j \leq n} |x_j - x_{ij}|. \tag{7}$$

Найдём минимальное и максимальное расстояние до «идеального» объекта:

$$l = \min_{1 \le i \le m} L_i, \quad (8)$$

$$L = \max_{1 \le i \le m} L_i. \quad (9)$$

Вычислим шаг

$$h = \frac{L - l}{p} \quad (10)$$

и распределим объекты по $p$ отрезкам $[l + (t-1)h, l + th]$, где $t = \overline{1, p}$.

Объект, попавший в отрезок $[l, l + h]$, будет иметь рейтинг $q_1 = 1$, а в отрезок $[l + (p-1), l + ph]$ – рейтинг $q_p = p$.

Рассмотрим тот же датасет. В качестве характеристик выберем общую площадь квартиры (в м$^2$), жилую площадь (в м$^2$), площадь кухни (в м$^2$), количество комнат, расстояние до центра (в км), год постройки, этаж, время пешего пути от дома до метро (в минутах). $m = 440,\ n = 8.$

Каждой числовой характеристике присвоим весовой коэффициент: для наиболее значимой $v_j = 8$, для наименее значимой $v_j = 1$ (подробное описание представлено в таблице 1, которая является авторской). Нормируем вектор характеристик по формуле (4).

Таблица 1 – Числовые характеристики и их весовые коэффициенты

| Характеристика | Весовой коэффициент $v_j$ | Нормированный весовой коэффициент $w_j$ |
|---|---|---|
| Общая площадь | 8 | $\frac{8}{36}$ |
| Жилая площадь | 7 | $\frac{7}{36}$ |
| Площадь кухни | 5 | $\frac{5}{36}$ |
| Кол-во комнат | 6 | $\frac{6}{36}$ |
| Расстояние до центра | 4 | $\frac{4}{36}$ |
| Время пути до метро | 1 | $\frac{1}{36}$ |
| Год постройки | 3 | $\frac{3}{36}$ |
| Этаж | 2 | $\frac{2}{36}$ |

Вычислим новые значения показателей по формуле (5) и нормируем матрицу по формуле (6). В каждом столбце матрицы выберем максимальный элемент, вектор *x*, соответствующий «идеальному» объекту, будет иметь все координаты, равные единице.

Оптимальное число кластеров было выбрано на основании суммы квадратов случайных ошибок, которая минимальна при трёх кластерах.

Для поиска расстояния до «идеального» объекта используем формулу (7). Воспользовавшись формулами (8)-(10), найдём $l = 0.65853047$, $L = 0.98076923$, $h = 0.10741292$.

Распределив объекты по отрезкам (см. таблицу 2, которая является авторской), получим, что в кластере «0» будет 53 квартиры, в кластере «1» – 198, в «2» – 182.

Таблица 2. – Рейтинг и границы кластеров

| Кластер | Рейтинг | Нижняя граница | Верхняя граница |
|---|---|---|---|
| «0» | $q_1 = 1$ | 0.61414791 | 0.73324146 |
| «1» | $q_2 = 2$ | 0.73324146 | 0.85233502 |
| «2» | $q_3 = 3$ | 0.85233502 | 0.97142857 |

Отметим полученные кластеры на плоскости. На рисунке 1, который является авторским, на горизонтальной оси отмечен год постройки дома, на вертикальной – количество комнат в квартире. Крестиками отмечены квартиры кластера «0», минусами – «1», звёздочками – «2».

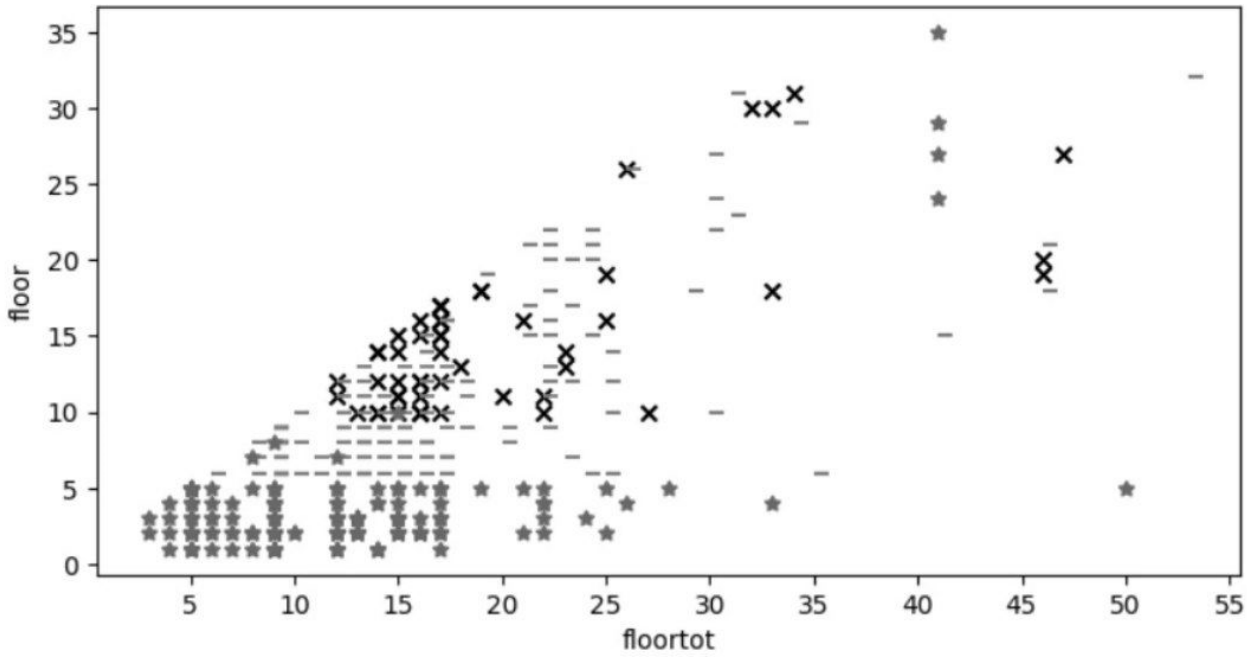


Рис. 1 – Распределение квартир по кластерам на плоскости этаж-этажность дома

Удалось заметить, что квартиры в пятиэтажках и квартиры в высотках до 5 этажа попадают в кластер «2». Квартиры на 8 и 9 этаже попадаюут в кластер «1». Все квартиры в домах до 1955 года постройки будут принадлежать кластеру «2». А объекты с одной или четырьмя комнатами не окажутся в кластере «0».

Чтобы выяснить, какой способ предсказания стоимости даёт наиболее точный результат, были найдены суммы квадратов случайных ошибок тренировочной и тестовой выборок. Данные представлены в таблице 3, которая является авторской.

Таблица 3. – Сравнение исследуемых моделей

| Название модели | Сумма квадратов случайных ошибок | |
|---|---|---|
| | для тренировочной выборки | для тестовой выборки |
| Регрессионная модель | $6.46 \cdot 10^{15}$ | $1.814 \cdot 10^{15}$ |
| Регрессионная модель после удаления незначимых коэффициентов | $6.48 \cdot 10^{15}$ | $1.832 \cdot 10^{15}$ |
| Регрессионная модель после устранения мультиколлинеарности | $6.55 \cdot 10^{15}$ | $1.761 \cdot 10^{15}$ |
| Регрессионная модель с переменной структурой | $5.7 \cdot 10^{15}$ | $1.506 \cdot 10^{15}$ |

| | | |
|---|---|---|
| Регрессионная модель с учётом рейтинга объектов | $5.26 \cdot 10^{15}$ | $1.528 \cdot 10^{15}$ |
| Регрессионная модель с переменной структурой с учётом рейтинга объектов | $4.56 \cdot 10^{15}$ | $1.397 \cdot 10^{15}$ |

Полученные результаты позволяют сделать следующие выводы: разбиение выборки на кластеры повышает точность регрессионной модели; устранение мультиколлинеарности может уменьшить случайные ошибки; регрессионная модель с переменной структурой с учётом рейтинга объектов даёт наилучший результат предсказания стоимости жилой недвижимости.

**Библиографический список:**

1. Сайт «Авито» [Электронный ресурс]. – Режим доступа – URL: https://www.avito.ru/moskva/nedvizhimost.

2. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. – 1022 с.

3. Богданова Т.К., Камалова А.Р., Кравченко Т.К., Полторак А.И. Проблемы моделирования оценки стоимости жилой недвижимости // Бизнес-информатика. – 2020. – Т.14. – №3. – С.7-23. DOI: 10.17323/2587-814X.2020.3.7.23.

4. Документация по Matplotlib [Электронный ресурс] – Режим доступа – URL: https://matplotlib.org/stable/index.html.

5. Кодкамп [Электронный ресурс] – Режим доступа – URL: https://www.codecamp.ru/blog/.

6. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.

7. Сигал И.Х., Иванова А.П. Введение в прикладное дискретное программирование: модели и вычислительные алгоритмы: Учеб. пособие. – 2-е изд., испр. и доп. – М.: ФИЗМАТЛИТ, 2007. – 304 с.

8. Стерник Г.М. Технология анализа рынка недвижимости. – М.: АКСВЕЛЛ, 2005. – 200 с.

9. Учебное пособие по парсингу сайтов на Python [Электронный ресурс] – Режим доступа – URL: https://bestprogrammer.ru/izuchenie/uchebnoe-posobie-po-parsingu-sajtov-na-python?ysclid=lgceo0tkl1695291175.
10. Эконометрика: учеб. / под ред. И.И. Елисеевой. – М. Проспект, 2009. – 288 с.

*Оригинальность 81%*