УДК 004 DOI 10.25730/VSU.0536.19.004

Сравнение классических регрессионных моделей с моделями, построенными с помощью продвинутых методов машинного обучения

А. В. Шатров[1], Д. Э. Пащенко[2]

[1]доктор физико-математических наук, профессор кафедры ЦТО, Вятский государственный университет. Россия, г. Киров. E-mail: avshatrov1@yandex.ru

[2]студент 4 курса ВятГУ, факультет компьютерных и физико-математических наук, кафедра ЦТО, Вятский государственный университет. Россия, г. Киров. E-mail: smile_dan@mail.ru

Аннотация. Эконометрика, или эконометрическое моделирование, занимается построением моделей и прогнозов, анализируя которые, можно понять текущее положение и направление развития экономики и ее отраслей. В данной работе рассматриваются классические эконометрические регрессионные модели и модели машинного обучения (Machine Learning) на основе использования современных методов и программных средств прикладной вычислительной статистики, проводится их построение, составление прогноза по ним, а также сравнение полученных результатов.

Ключевые слова: эконометрика, эконометрическое моделирование, регрессия, машинное обучение.

Введение

В современном мире важное место занимает планирование и прогнозирование, с которым мы можем столкнуться повсеместно. Наиболее важна их роль, разумеется, в экономике. Тем не менее современная экономика должна не только отвечать на самые привычные вопросы (поиск ресурсов, производство и сбыт продукции), но и иметь представление о развитии какого-либо рынка в частности или всей экономики в целом. Методы анализа данных и прогнозирования с помощью современных программных сред являются в настоящее время наиболее востребованными с точки зрения приложений в математической статистике. Задачи анализа данных относятся к разделам прикладной вычислительной статистики и востребованы практически во всех отраслях современной науки. В ходе работы были изучены теоретические основы анализа и прогнозирования, построены модели по общедоступным публикуемым данным о стоимости жилья. Данные опубликованы ПАО «Сбербанк» на сайте Kaggle [1]. Построение всех моделей и работа с данными проведена с помощью языка программирования Python [2]. Перед построением моделей данные были стандартизированы (с помощью функции *StandardScaler* библиотеки *sklearn*), цены на квартиры – продефлированы, так как период наблюдения составлял несколько лет.

Постановка задачи

Ввиду того что данные, предоставленные сайтом, содержат большое число факторов, необходимо было использовать модель множественной линейной регрессии.

Уравнение множественной линейной регрессии имеет вид

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e,$$

где y – зависимая переменная, $x_1, \dots, x_n$ – независимые или объясняющие переменные; e – стохастическая переменная, включающая влияние неучтенных факторов. Параметры $b_1, \dots, b_n$ – коэффициенты регрессии, характеризующие среднее изменение результата с изменением соответствующего фактора на единицу при неизменном значении других факторов, закрепленных на среднем уровне [3].

Для оценки построенной модели используют следующие показатели:

1. Коэффициент детерминации: $R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$, который показывает взаимосвязь между переменными. Коэффициент лежит в промежутке $[0; 1]$. Чем ближе его значение к единице, тем более точной является модель.

2. Среднеквадратичная ошибка: $MSE = \frac{1}{n}\sum(y - \hat{y})^2$.

3. Средняя абсолютная ошибка: $MAE = \frac{1}{n}\sum|y - \hat{y}|$.

В случае с ошибками меньшее значение является лучшим.

4. Коэффициент несоответствия Тейла: $K = \sqrt{\frac{\sum(y_t - \hat{y}_t)^2}{\sum y_t^2 + \sum \hat{y}_t{}^2}}$. Его значение принадлежит промежутку от 0 до 1. Модель считается очень хорошей, если значение находится в диапазоне от 10% до 15% ($\hat{y}$ – предполагаемое, или предсказанное, значение; y – реальное значение).

Для оценки качества модели также используются и другие показатели.

Построение моделей

Данные были разделены на две части. Первая выборка была обучающей, она составляла примерно 80% от общего числа данных. На ее примере мы и строили как модель множественной линейной регрессии, так и модель градиентного бустинга. Вторая выборка – тестирующая, или проверочная, – составляла оставшиеся 20% данных. По ней проводилась оценка качества модели путем сравнения этих данных с прогнозом, сделанным по построенной модели.

При работе с данными и их обработке были выделены 30 наиболее влиятельных факторов. Для этого была применена функция *feature importance* библиотеки *sklearn* языка *Python* [2]. К ним, разумеется, были отнесены все независимые факторы, указанные выше. Далее были проанализированы их значимости (*p*-показатель значимости), проведен шаговый регрессионный анализ, исключающий факторы. Отбор проводился по *F*-критерию.

Суть метода включения – в последовательном исключении переменных из модели до тех пор, пока регрессионная модель не будет ухудшаться от исключения очередного фактора.

В итоге мы получили модель, которая содержала 22 фактора:

$$y = 0.6653\,a + 0.1804\,b - 0.1604\,c - 0.1981\,d + 0.1093\,e + \\ + 0.0904\,f + 0.2076\,g + 0.0083\,h + 0.0899\,i - 0.0853\,j - \\ - 0.0249\,k + 0.0199\,l - 0.0425\,m - 0.0093\,n - 0.1298\,o - \\ - 0.0697\,p - 0.0196\,q - 0.0118\,r + 0.0083\,s - 0.0855\,t + \\ + 0.0275\,u + 0.0736\,v$$

Интерпретация переменных: *a* – площадь квартиры, *b* – жилая площадь, *c* – рейтинг района, *d* – полезная площадь кухни, *e* – этаж, *f* – количество комнат, *g* – площадь кухни, *h* – средний размер комнат, *i* – высота квартиры, *j* – близость к учреждениям культуры, *k* – этажей в доме, *l* – возраст дома, *m* – близость к точкам общепита, *n* – количество университетов в округе, *o* – близость к учреждениям спорта, *p* – близость к шоссе, *q* – район, *r* – близость к паркам, *s* – средняя полезная площадь комнат, *t* – близость к остановкам общественного транспорта, *u* – рейтинг этажа, *v* – близость к промышленности.

Модель адекватна и имеет достаточно неплохие показатели качества (таблица 1), полученные с помощью кросс-валидации (перекрестной проверки).

Таблица 1

Показатели качества модели множественной линейной регрессии

R^2	*MSE*	*RMSE*	*MAE*	*F*-статистика	Коэффициент Тейла
0,507	0,5	0,707	0,415	1199	0,41

По данным таблицы видно, что мы имеем сравнительно небольшую среднеквадратичную ошибку (*MSE*) и среднюю абсолютную ошибку (*MAE*), очень хорошее значение *F*-статистики. Однако коэффициент детерминации (R^2), а также коэффициент несоответствия Тейла показывают, что полученная модель хоть и объясняет большую часть данных, но прогноз, сделанный по ней, вполне может оказаться неточным.

Графики распределения остатков, сравнения теоретического и реального распределения остатков также покажут вполне приемлемый результат, который подтвердит, что модель достаточно неплоха.

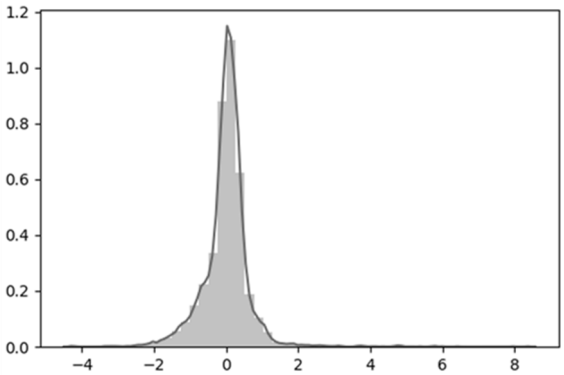

Рис. 1. График распределения остатков предсказания и реальных данных

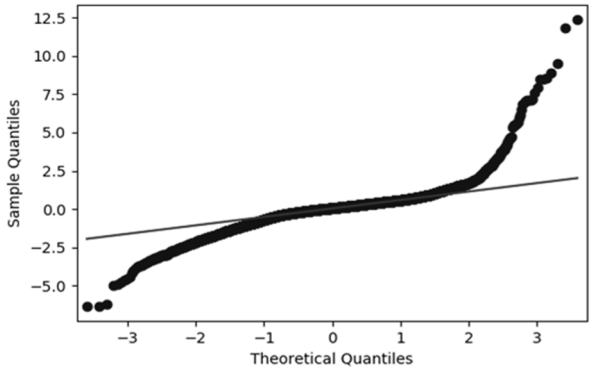

Рис. 2. График сравнения теоретического и реального распределения остатков

Если сравнение реальных данных с результатами прогноза по модели представить в виде графика, построенного в координатах цены (зависимого фактора) и полной площади квартиры (как самого ключевого независимого фактора), то мы получим достаточно неплохое совпадение, хотя и далекое от идеального, что подтверждают найденные нами ранее показатели качества модели (R^2 недостаточно высокий, коэффициент Тейла, напротив, высок).

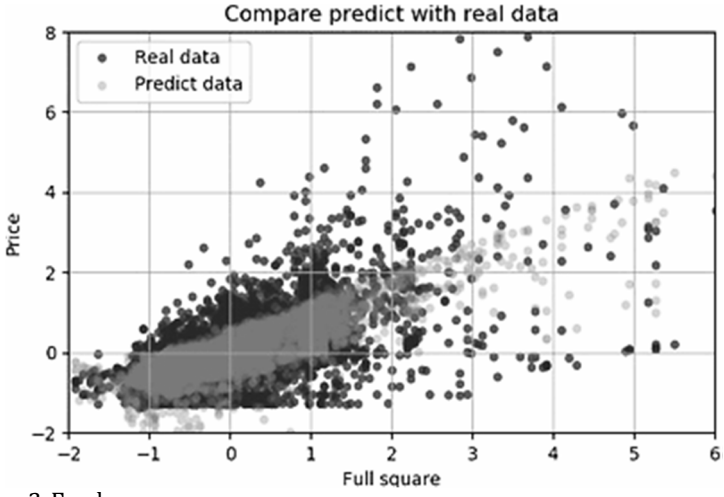

Рис. 3. График сравнения спрогнозированных данных с реальными

Еще один показатель качества модели – отсутствие мультиколлинеарности. Проверим модель, построив матрицу парных корреляций. По ней мы увидим, что не наблюдается высоких парных коэффициентов корреляции. Хотя есть и высокие значения, однако ни одно из них не доходит до значения 0,7, которое является критическим для существования мультиколлинеарности.

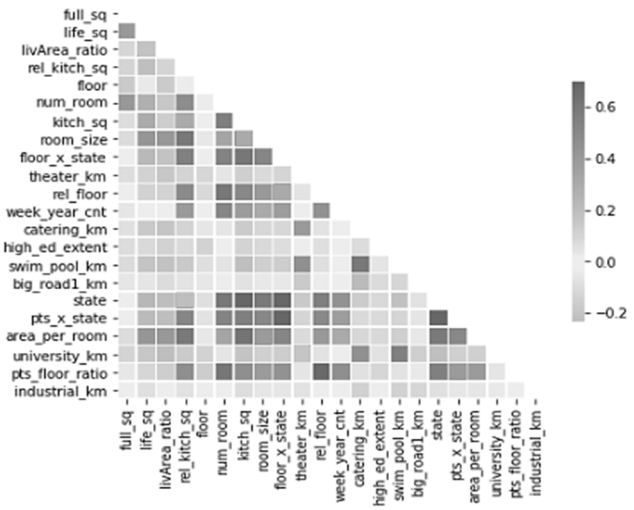

Рис. 4. Матрица парных корреляций

Можно сделать вывод, что данная модель множественной линейной регрессии вполне применима и корректна, хоть и имеет некоторые показатели, далекие от идеальных.

Далее была рассмотрена модель градиентного бустинга *lightGBM*. Градиентный бустинг – метод машинного обучения, применяемый для задач регрессии и классификации. Он создает модель прогнозирования в виде ансамблей слабых моделей прогнозирования, обычно деревьев решений; строит модель поэтапно, обобщает их, позволяя оптимизировать произвольную дифференцируемую функцию потерь [4].

Полученные результаты приведены в таблице 2.

Таблица 2

Показатели качества модели градиентного бустинга

R^2	*MSE*	*RMSE*	*MAE*
0,734	0,14	0,375	0,277

Получаем, что модель намного лучше, чем построенная нами ранее. Показатели всех ошибок значительно ниже, коэффициент детерминации объясняет практически ¾ данных, что больше на четверть, чем модель множественной линейной регрессии.

Полученные результаты достаточно хорошие, однако значение среднеквадратической ошибки можно постараться снизить. Для этого попробуем использовать автоматический подбор параметров для модели с помощью функции *GridSearch* (поиск по сетке). Данная функция перебирает все возможные комбинации параметров, которые мы задали, а затем выводит наиболее хороший вариант. Ввиду того что у функции *lightGBM* 7 параметров и у каждого из них по 3 предложенных значения, функции придется перебирать большое число вариантов. В итоге данная процедура выполнялась около пяти часов (расчет происходил на компьютере с двухъядерным процессором, имеющим четыре потока выполнения и максимальную тактовую частоту в 2,8 ГГц). Процедуру поиска по сетке можно проводить неоднократно, не забывая изменить параметры, это может улучшить модель.

Подставив полученные параметры и построив модель повторно, мы получим следующие результаты (таблица 3).

Таблица 3

Показатели качества модели градиентного бустинга с подобранными параметрами

R^2	*MSE*	*RMSE*	*MAE*
0,749	0,092	0,303	0,274

Проанализируем значения. Мы видим, что показатели ошибок лучше у данной модели. Этот факт достаточно закономерен, так как наша функция поиска по сетке *GridSearch* оптимизировала модель именно по критерию *RMSE*. Коэффициент детерминации также чуть больше у данной модели. В таблице 4 сравниваются показатели качества всех построенных моделей

Таблица 4

Показатели качества построенных моделей

Модель	R^2	*MSE*	*RMSE*	*MAE*
Множественная линейная регрессия	0,507	0,5	0,707	0,415
Градиентный бустинг (1)	0,734	0,14	0,375	0,277
Градиентный бустинг (2)	0,749	0,092	0,303	0,274

В итоге мы получаем, что лучше вторая модель *lightGBM*, построенная по результатам *GridSearch*.

Вывод

По результатам построения моделей мы можем говорить, что самым значимым фактором ценообразования квартир является ее площадь. Также существует порядка 5-10 чуть менее значимых факторов.

Модель градиентного бустинга дает более хорошие результаты, особенно если подобрать параметры функции поиском по сетке, однако ее построение требует большего объема знаний, а также определенных затрат по времени (если выполнять поиск по сетке для оптимизации модели).

Следовательно, достаточно простой моделью, но при этом емкой и сравнительно не искажающей данные, будет являться модель множественной линейной регрессии, построенная по 22 факторам.

Стоит отметить, что современные возможности машинного обучения позволяют упростить расчеты, улучшить модель и сэкономить время. Их также следует использовать в анализе и прогнозировании данных.

Список литературы

1. Description of LightGBM. URL: https://lightgbm.readthedocs.io/en/latest/Parameters.html (дата обращения: 28.11.2018).
2. Sberbank Russian Housing Market. URL: https://www.kaggle.com/c/sberbank-russian-housing-market.html (дата обращения: 31.10.2018).
3. *Плас Дж. Вандер* Python для сложных задач: наука о данных и машинное обучение. СПб. : Питер, 2018. 576 с.
4. Эконометрика : учебник для бакалавриата и магистратуры / И. И. Елисеева [и др.]; под ред. И. И. Елисеевой. М. : Юрайт, 2016. 449 с.

Comparison of classical regression models with models built using advanced machine learning methods

A. V. Shatrov[1], D. E. Paschenko[2]

[1]doctor of physical and mathematical sciences, professor of the Department of digital technologies in education, Vyatka State University. Russia, Kirov. E-mail: avshatrov1@yandex.ru

[2]student of the 4 course of VyatSU, Faculty of computer and physical and mathematical sciences, Department of digital technologies in education, Vyatka State University. Russia, Kirov. E-mail: smile_dan@mail.ru

Abstract. Econometrics, or econometric modeling, is engaged in the construction of models and forecasts. Analyzing that, you can understand the current situation and direction of development of the economy and its industries. This paper discusses the classical econometric regression models and machine learning models (Machine Learning) based on the use of modern methods and software of applied computational statistics, their construction, forecasting on them, as well as comparison of the results.

Keywords: econometrics, econometric modeling, regression, machine learning.

References

1. Description of LightGBM. Available at: https://lightgbm.readthedocs.io/en/latest/Parameters.html (date: 28.11.2018).
2. SberbankRussianHousingMarket. Available at: https://www.kaggle.com/c/sberbank-russian-housing-market.html (date: 31.10.2018).
3. *Place John Wander Python dlya slozhnyh zadach: nauka o dannyh i mashinnoe obuchenie* [Python for complex tasks: data science and machine learning]. SPb. Piter. 2018. 576 p.
4. *Ekonometrika : uchebnik dlya bakalavriata i magistratury* – Econometrics: textbook for undergraduate and master students / I. I. Eliseeva [et al.]; ed. I.I. Eliseeva. M. Yurayt. 2016. 449 p.