УДК 004.9

Сравнительный анализ результатов машинного обучения и регрессионной модели траекторий поведения пользователей онлайн-сервисов

Е.А. Шипилова, Е.Е. Некрылов

Воронежский государственный университет, Воронеж, Россия

Динамика поведения интернет-покупателей вызывает большой интерес у маркетологов с целью максимизации прибыли магазина и прогнозирования развития онлайн-продаж. Наиболее широко распространенными методами обработки статистических данных являются методы регрессионного статистического анализа, кроме того, актуальность приобретают методы машинного обучения. Цель исследования заключалась в прогнозировании поведения пользователей онлайн-магазина, на основе исходных данных, получаемых технологиями BigData. Проведена оценка корреляции фактора и результата, показано наличие прямой линейной зависимости. Классическими методами регрессионного анализа были определены коэффициенты линейной регрессии, оценена их значимость, адекватность модели, ошибки аппроксимации. Методами машинного обучения обучена модель, определены коэффициенты. Сравнительные результаты представлены в виде графика. Определен доверительный интервал прогноза для уровня значимости $\alpha = 0,05$. Представлены соответствующие выводы.

Ключевые слова: регрессионный анализ, машинное обучение, коэффициент корреляции, коэффициенты регрессии, адекватность модели, доверительный интервал прогноза.

Comparative analysis of machine learning results and regression model of online service user behavior trajectories

E.A. Shipilova, E.E. Nekrylov

Voronezh State University, Voronezh, Russia

The dynamics of behavior of online customers is of great interest to marketers in order to maximize the profit of the store and predict the development of online sales. The most widespread methods of processing statistical data are regression statistical analysis methods, and machine learning methods acquire relevance. The purpose of the study was to predict the behavior of online store users, based on the original data obtained by BigData technologies. The correlation of the factor and the result was evaluated, the presence of a direct linear relationship was shown. Classical regression analysis methods would determine linear regression coefficients, assess their significance, model adequacy, mean absolute and relative approximation errors. The model was trained by machine learning methods, coefficients were determined. Comparative results are presented in the form of a graph. The prediction confidence interval was determined for the significance level $\alpha = 0,05$. Relevant findings are presented.

Keywords: regression analysis, machine learning, correlation coefficient, regression coefficients, model adequacy, prediction confidence interval.

Целью пользователей глобальной сети Интернет все чаще является поиск товаров для их приобретения, в связи с этим, в настоящее время широкое распространение получают онлайн-магазины или сайты продаж. Динамика поведения интернет-покупателей вызывает большой интерес у маркетологов [1] с целью максимизации прибыли магазина и прогнозирования развития онлайн-продаж. Для успешной интернет-торгов-

© Шипилова Е.А., Некрылов Е.Е., 2023

ли и продвижения онлайн-магазинов достаточно остро стоит вопрос прогнозирования количества покупок тех или иных товаров, а также пользователей онлайн-сервисов для своевременного укомплектования складов необходимым ассортиментом [2, 3]. Наиболее широко распространенными методами обработки статистических данных на сегодняшний день являются методы регрессионного статистического анализа, кроме того актуальность приобретают методы машинного обучения [4].

Цель исследования заключалась в прогнозировании поведения пользователей онлайн-магазина, для чего, на основе технологии BigData были получены исходные данные, которые после необходимой предварительной сортировки и обработки можно представить в следующем виде (рис. 1).

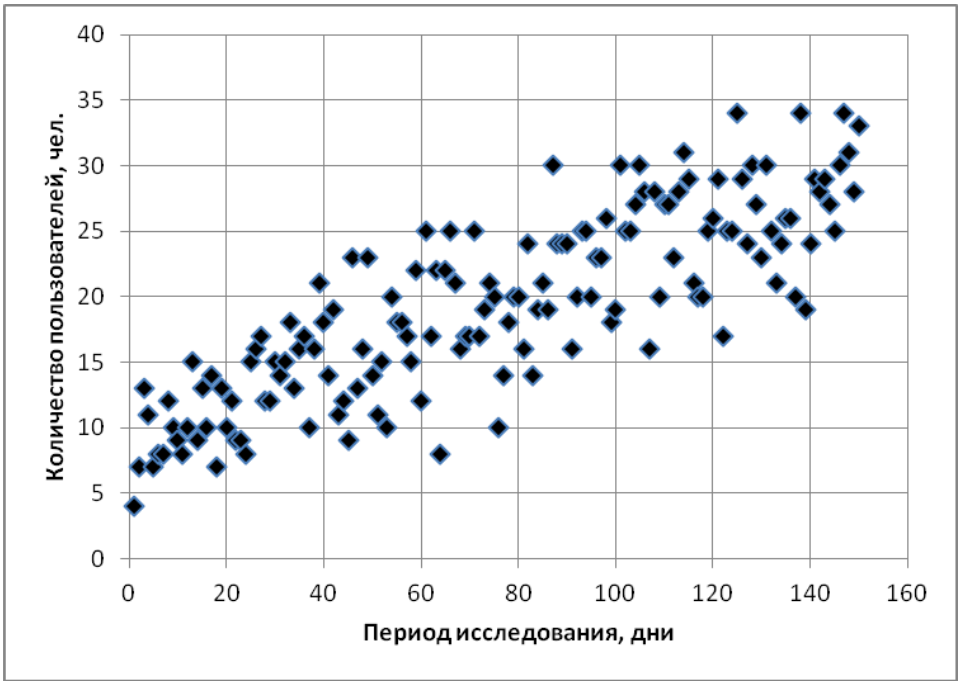

Рисунок 1. Исходные данные.

Анализируя полученные данные (рис. 1), можно предположить, что между фактором (текущий день периода исследования) и результатом (количество пользователей) существует зависимость. Также из рис. 1 можно предположить линейный вид рассматриваемой зависимости. Уравнение линейной регрессии при этом можно записать в виде: $y^m = a + b \cdot x$, где x – значение фактора (день), y^m – модельное значение результата (количество пользователей), a, b – оценки параметров модели. Оценим тесноту связи фактора и результата. Линейный коэффициент парной корреляции:

$$r_{xy} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0,832 \quad (1)$$

Согласно шкале Чеддока, между показателем y и фактором x существует высокая связь.

Коэффициент детерминации: $R^2 = r_{xy}^2 = 0,832^2 = 0,692$; что означает, вариация результата y (количество пользователей) на 69,2 % объясняется вариацией фактора *x* (длительностью работы сервиса).

Используя известный метод наименьших квадратов, оценки параметров линейной модели *a* и *b* можно рассчитать по формулам:

$$a = \bar{y} - b \cdot \bar{x}, \quad (2)$$

$$b = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}, \quad (3)$$

где числитель в (2) – ковариация признаков регрессионной модели, знаменатель – дисперсия фактора *x* (σ_x^2).

Проведя необходимые расчеты, были получены следующие оценки параметров модели: *a* = 9,209; *b* = 0,135. Полученная зависимость представлена на рис. 2.

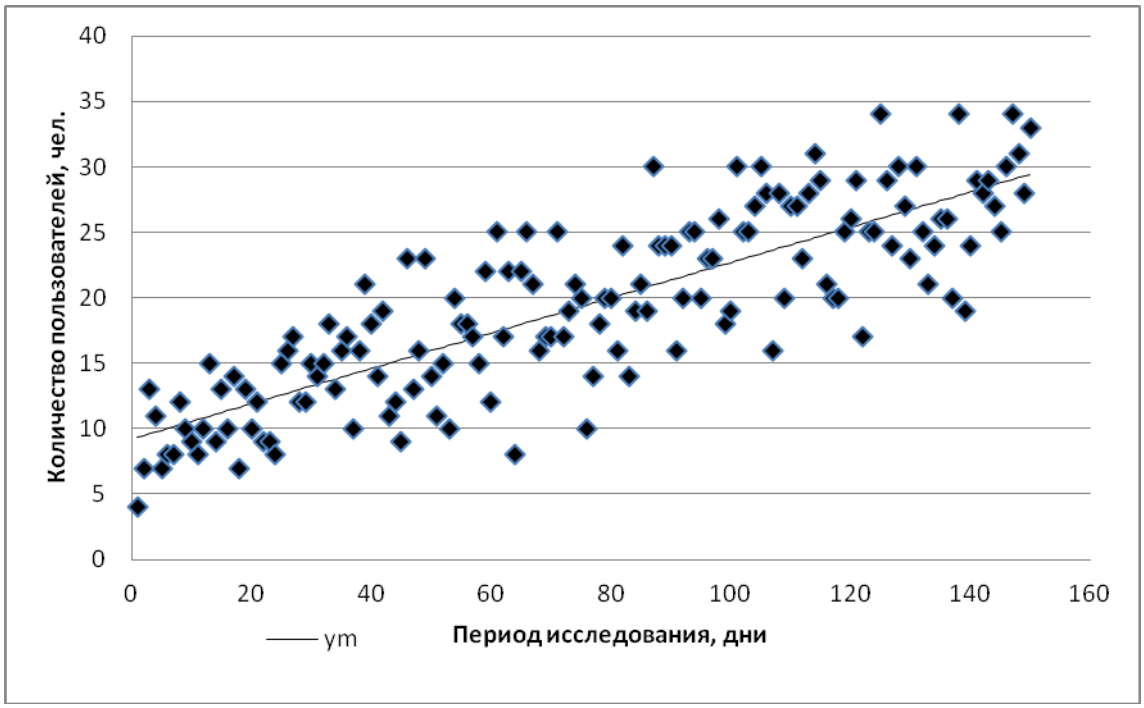

Рисунок 2. Исходные данные и регрессионная модель.

Оценим значимость коэффициентов, адекватность модели (статистическую надежность результатов регрессионного моделирования). Оценка значимости модели с помощью критерия Стьюдента проводится путем сравнения их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \qquad t_a = \frac{a}{m_a};$$

$$m_a = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^m\right)^2}{(n-2)} \cdot \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = S_{\text{ост}} \frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{n\sigma_x} = 0,644;$$

$$m_b = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^m\right)^2}{(n-2)\cdot\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = \frac{S_{ост}}{\sigma_x\sqrt{n}} = 0,007;$$

$$t_b = 18,222; \qquad t_a = 14,294.$$

Табличное значение критерия Стьюдента определяется в зависимости от числа степеней свободы ($n - m - 1$) и для уровня значимости $\alpha = 0,05$: $t_{кр} = 1,97$. Сравнивая фактические и табличное значения критерия Стьюдента $t_a > t_{кр}$, $t_b > t_{кр}$ можно сделать вывод о значимости модели по параметрам.

Оценку значимости уравнения регрессии проводили с помощью F-критерия Фишера:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2}\cdot(n-2) = 332,048.$$

Табличное значение критерия определяется в зависимости от параметров $k_1 = m$, параметра k_2 – числа степеней свободы ($n - m - 1$), для уровня значимости $\alpha = 0,05$ $F_{кр} = 3,9$. Так как $F > F_{кр}$, уравнение регрессии статистически значимое.

Средняя относительная ошибка аппроксимации:

$$\bar{A} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left|\frac{y_i - y_i^m}{y_i}\right|\cdot 100\% = 19,68\%.$$

В среднем расчетные значения y_m для линейной модели отличаются от фактических значений на 19,68%, что для подобных объектов допустимо.

Средняя квадратичная ошибка регрессионной модели:

$$\sigma_y^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - y_i^m\right)^2 = 15,205.$$

Средняя абсолютная ошибка:

$$\bar{E} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left|y_i - y_i^m\right| = 3,186.$$

Машинное обучение модели проводилось с использованием линейной модели на 80% исходных данных. Проведя необходимые расчеты, были получены следующие оценки параметров линейной модели: $a = 8,794$; $b = 0,144$. Рассчитанная методом машинного обучения зависимость в сравнении с регрессионной моделью представлена на рис. 3. Необходимо проверить качество полученной модели на оставшихся 20% данных.

Средняя квадратичная ошибка:

$$\sigma_y^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - y_i^m\right)^2 = 18,92.$$

Средняя абсолютная ошибка:

$$\bar{E} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left|y_i - y_i^m\right| = 3,468.$$

Рассчитаем прогнозное значение результата, если прогнозное значение фактора увеличится на 10% от значения его последнего уровня, т.е. для 165 дня.

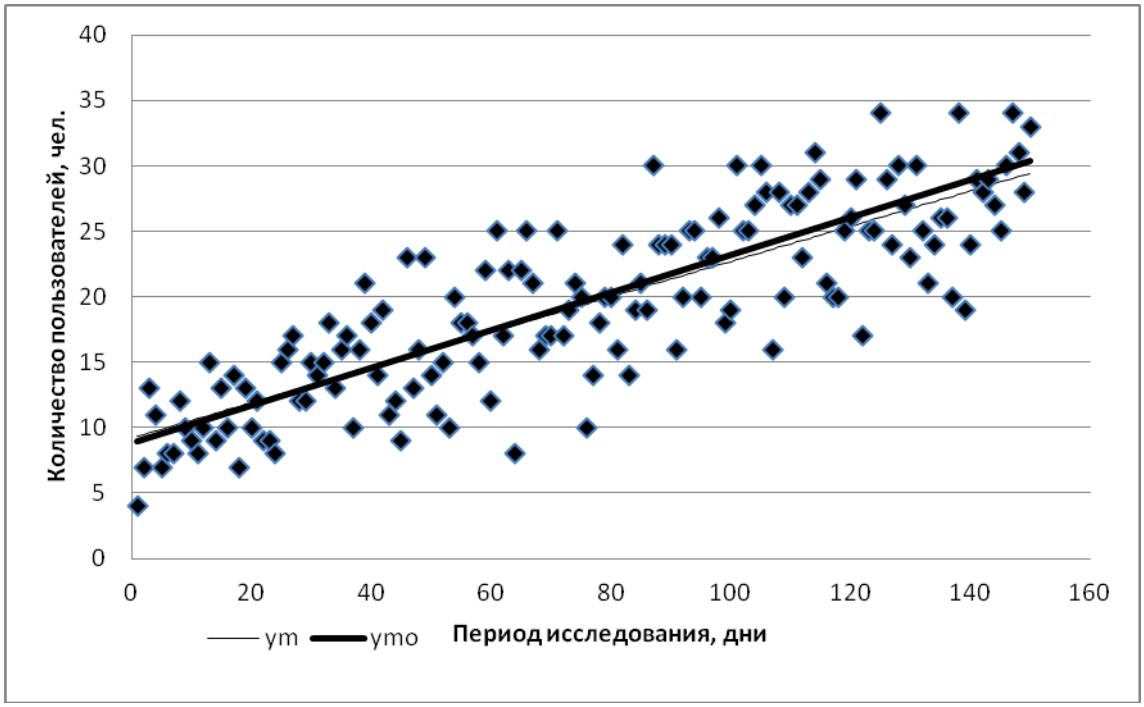

Рисунок 3. Исходные данные, регрессионная модель и модель машинного обучения.

Прогнозное значение результативного признака (количество пользователей интернет-магазина) определим по линейной модели подставив в него планируемую величину фактора x – день:

$$y^m = 9,209 + 0,135 \cdot x .$$

Среднее значение фактора x составляет 75,5. Если его увеличить на 10% $\bar{x}$ составит 83,05, тогда количество пользователей составит: $y^m = 9,209 + 0,135 \cdot x = 20,421$.

Определим доверительный интервал прогноза для уровня значимости $\alpha = 0,05$.

Средняя стандартная ошибка прогноза:

$$m_{\bar{y}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^m\right)^2}{n-m-1}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(\hat{x} - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = 3,94 .$$

Средняя ошибка прогноза (табличное значение критерия Стьюдента $t_{табл} = 1,97$)

$$\Delta\bar{\varepsilon} = t_{табл} \cdot m_{\bar{y}} = 7,76 .$$

Строим доверительный интервал прогноза:

$$\gamma_{\bar{y}} = y^m \pm \Delta\bar{\varepsilon} = 20,421 \pm 7,76 .$$

То есть, с вероятностью 0,95 можно утверждать, что по истечении 10% от рассмотренной длительности работы сервиса, значение количества пользователей будет заключено в пределах от 13 до 28 человек.

По полученным результатам можно сделать следующие выводы:

Между фактором *x* и результатом *y* существует сильная и прямая корреляционная зависимость. Качество линейной модели оценивается как удовлетворительное, т.к. средняя ошибка аппроксимации для нее превышает 20%. Модель можно признать адекватной и реальной экономической ситуации. Классические методы регрессионного анализа дают более точные результаты, по сравнению с методами машинного обучения.

Выполненный прогноз достаточно неточен, так как диапазон верхней и нижней границ доверительного интервала различается почти в два раза.

СПИСОК ИСТОЧНИКОВ

1. Digital 2021: главная статистика по России и всему миру [Электронный ресурс]. – URL: https://spark.ru/user/115680/blog/74085/digital2021-glavnaya-statistika-po-rossii-i-vsemu-miru/ (дата обращения: 15.03.2023).

2. Прохорова М.В. Организация работы интернет-магазина: Пособие / М.В. Прохорова, А.Л. Коданина. – 3-е изд. М.: Дашков и К, 2020. – 332 с.

3. Gull M. Customer Behavior Analysis Towards Online Shopping using Data Mining, / M. Gull, A. Pervaiz // 2018 5th International Multi-Topic ICT Conference (IMTIC). – 2018. – pp. 1-5.

4. Шипилова Е.А. Анализ и моделирование траекторий поведения пользователей онлайн-сервисов с использованием платформы RETENTIONEERING / Е.А. Шипилова, Е.Е. Некрылов, Т.В Курчекова // Моделирование систем и процессов: Научно-технический журнал. – 2022. – Т. 15. – Вып. 4. – С. 82-93.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Шипилова Елена Алексеевна, кандидат технических наук, доцент, доцент кафедры теории функции и геометрии, Воронежский государственный университет, Воронеж, Россия.

Некрылов Егор Евгеньевич, студент 5-го курса, Воронежский государственный университет, Воронеж, Россия.