Моделирование и анализ данных
2020. Том 10. № 4. С. 41–50
DOI: https://doi.org/10.17759/mda.2020100404
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2020. Vol 10, no. 4, pp. 41–50
DOI: https://doi.org/10.17759/mda.2020100404
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.85

Использование методов машинного обучения для решения задач прогнозирования спроса на новый товар в интернет-маркетплейсе

Осин А.А.*
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-2664-1370
e-mail: artemosin1@yandex.ru

Фомин А.К.**
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-3545-4435
e-mail: artem.fomin@outlook.com

Сологуб Г.Б.****
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

Виноградов В.И.******
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-3773-9653
e-mail: vvinogradov@inbox.ru

Для цитаты:
Осин А.А., Фомин А.К., Сологуб Г.Б, Виноградов В.И. Использование методов машинного обучения для решения задач прогнозирования спроса на новый товар в интернет-маркетплейсе // Моделирование и анализ данных. 2020. Том 10. № 4. С. 41–50. DOI: https://doi.org/10.17759/mda.2020100404

***Осин Артем Александрович,** Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-2664-1370, e-mail: artemosin1@yandex.ru

****Фомин Артем Константинович,** Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-3545-4435, e-mail: artem.fomin@outlook.com

Работа направлена на исследование возможности применения методов машинного обучения для построения моделей прогнозирования спроса на новые товары в интернет-магазине Ozon.ru. Предлагаются к рассмотрению ранее неиспользуемые в рамках конкретной задачи подходы к решению. В качестве выборки использованы данные об истории продаж и хранении товаров на сайте Ozon.ru. Приводится описание, анализ примерного убытка сайта Ozon.ru, используемых данных, процесса построения базовой модели, а также полученных результатов. Описываются метрики, использованные для оценки результатов прогнозирования, а также проводится сравнительный анализ между результатами предсказания построенной модели и результатами эвристически подобранных значений.

Ключевые слова: прогнозирования спроса, новый товар, энкодинг, градиентный бустинг, регрессия, препроцессинг, обработка данных, машинное обучение.

****Сологуб Глеб Борисович,*** кандидат физико-математических наук, доцент кафедры, Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

*****Виноградов Владимир Иванович,*** кандидат физико-математических наук, доцент кафедры, Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-3773-9653, e-mail: vvinogradov@inbox.ru

1. ВВЕДЕНИЕ

Всё больше участников рыночной экономики стремятся предоставлять свои товары и услуги онлайн. Поэтому неудивительно, что крайнюю востребованность приобрели интернет-маркетплейсы, которые выступают посредниками между предпринимателями и конечным потребителем.

Маркетплейсы предоставляют услуги предпринимателям, связанные с распространением их продукции, а также имеют возможность предоставлять потребителям свои услуги и товары. И в том, и в другом случае крайне актуальна задача предсказания спроса, так как и у маркетплейса и у предпринимателей, пользующихся услугами маркетплейсов, появляется возможность эффективно управлять денежными ресурсами – закупать только те товары, которые будут хорошо продаваться. То есть и сама платформа, и поставщик смогут лучше реагировать на изменения рынка и получать большую прибыль.

Цель работы заключается в проектировании системы для прогнозирования спроса на новые товары в интернет-маркетплейсе. В частности, необходимо разработать отдельный модуль, задача которого прогнозировать продажи товаров-новинок в зависимости от описательных характеристик и внешней информации.

Сформулируем этапы решения этой задачи:

1. Вычисления примерного убытка интернет-маркетплейса «Озон» для возможности интерпретации результата работы

2. Разработка базовой модели, которая работает с некоторым небольшим количеством признаков каждого товара и позволит оценить качество результатов работы конечной модели, а также необходима для понимания предоставленных данных.
3. Разработка специализированных моделей, для которых будут различаться природа исходных данных, а именно описание товара, изображения товара и данные, полученные из внешних источников.
4. Построение комбинированной модели на основе специализированных моделей.

2. ОПИСАНИЕ ДАТАСЕТА И ПРЕДОБРАБОТКА

Для построения базовой модели были представлены данные интернет-маркетплейсом «Озон», в частности, история продаж товаров, размером 1886147 уникальных продаж, история товаров на складе размером 13632607 уникальных значений и информация о категориях товаров размером 483280 значений.

	categorylevel1	fake_itemid
0	TV, audio, Hi-Fi & other electronics	0
1	TV, audio, Hi-Fi & other electronics	1
2	TV, audio, Hi-Fi & other electronics	2
3	TV, audio, Hi-Fi & other electronics	3
4	TV, audio, Hi-Fi & other electronics	4

Рис. 1. Информация о категориях товара

Информация о категориях включает в себя категорию и уникальный идентификатор товара.

	date	qty	fake_itemid
0	2015-01-05	1	94489280563
1	2015-03-03	1	94489280563
2	2016-10-10	1	94489280563
3	2016-05-19	1	94489280563
4	2016-11-21	1	94489280563
...	...	...	...
1886142	2020-07-22	1	171798711788
1886143	2020-07-22	1	51539627534
1886144	2020-07-20	1	51539627534
1886145	2020-07-22	1	137438973480
1886146	2020-07-20	1	85899365801

Рис. 2. Вид истории товара на складе

История продаж включала в себя дату продажи, количество и уникальный идентификатор товара.

	moment	price	itemdiscount	virtualdiscount	freeqty	fake_itemid
2077	2019-09-01 00:00:00.0	100.0	5.000000	5.000000	11	111669149762
2108	2019-06-06 00:00:00.0	100.0	13.000000	13.000000	14	111669149762
2110	2019-06-11 00:00:00.0	100.0	13.000000	13.000000	9	111669149762
2112	2019-06-12 00:00:00.0	100.0	13.000000	13.000000	8	111669149762
2122	2018-09-03 00:00:00.0	100.0	10.000000	10.000000	6	111669149762
...	...	...	...	...	...	...
13632591	2020-07-25 00:00:00.0	2200.0	50.062500	50.062500	0	171798712100
13632596	2020-07-25 00:00:00.0	800.0	30.125000	30.125000	0	68719496927
13632598	2020-07-25 00:00:00.0	430.0	9.773438	9.773438	0	85899366074
13632600	2020-07-25 00:00:00.0	4086.0	33.343750	33.343750	0	111669169791
13632604	2020-07-25 00:00:00.0	13455.0	23.078125	23.078125	0	111669169794

Рис. 3. Вид истории продаж товара

Перед обучением модели была произведена первичная обработка данных. В данных были убраны аномалии и пропущенные значения. Затем было необходимо обработать категориальные признаки. В контексте решаемой задачи было необходимо преобразовать категорию товара из текстового формата в целочисленный. Для обработки категориальных признаков существуют два основных подхода – one hot encoding [1] и label encoding [2].

Оба алгоритма работают схожим образом – среди обрабатываемых значений находятся уникальные и из этих значений создаются классы для будущего энкодера. Затем каждому значению выдается метка, означающая принадлежность к тому или иному классу. Различие этих двух методов заключается в способе разметки данных на классы. В случае с label encoding метка класса представляет собой целое число, принадлежащее интервалу $(1,...,n-1)$ где n – число уникальных значений признака. То есть результатом преобразования является столбец, содержащий в себе целые числа. One hot encoder, в свою очередь, создает $(1,...,n-1)$ столбцов, содержащих в себе либо нуль, либо единицу. Каждый столбец представляет собой выделенный уникальный класс. Единица в столбце означает принадлежность к классу, нуль – не принадлежность, соответственно.

Поскольку применение подхода one hot encoding приводит к увеличению размерности матрицы признаков и делает матрицу более разреженной, что может негативно сказаться на результатах обучения модели, нами было принято решение использовать label encoding. Предобработанные данные были соединены в итоговый датасет.

Из обработанных ранее данных были выделены следующие признаки:

- Количество конкурирующих товаров внутри группы категориального дерева;
- Цена товара;
- Категория, к которой относится товар;
- Средняя цена в категориальной группе, содержащей товар, на момент появления товара;
- Средние продажи в категориальной группе в прошлом месяце;
- Средние продажи в категориальной группе три месяца назад;
- Средние продажи в категориальной группе шесть месяцев назад;
- Средние продажи в категориальной группе год назад.

После извлечения описанных признаков, в датасете были оставлены только данные о первом появления товара на торговой площадке. В качестве целевого признака для обучаемой модели было взято количество продаж товара за текущий месяц. Полученный датасет был разбит на обучающую и тестовую выборку в соотношении 80/20.

3. ОПИСАНИЕ МОДЕЛИ

В качестве модели была использована LightGBM [3]. Это быстрый, распределенный, высокопроизводительный градиентный бустинг, основанный на деревьях решений. Он часто используется для задач, классификации, регрессии, ранжирования и других задач машинного обучения. Он обладает несколькими преимуществами перед другими градиентными бустингами [6]:

1. Быстрая скорость обучения и высокую эффективность за счет использования подхода роста деревьев в глубину
2. Низкое потребление памяти.
3. Более высокая точность, чем у других алгоритмов градиентного бустинга за счет построения более сложной структуры решающих деревьев. Однако, иногда это может привести к переобучению модели.
4. Совместимость с большими наборами данных. Способен также хорошо работать с большими наборами данных со значительным сокращением времени обучения по сравнению с XGBOOST.

Для оценки качества предсказаний модели, было принято решение использовать следующие метрики:

Средний модуль отклонения (MAE – Mean Absolute Error или MAD – Mean Absolute Deviation):

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |,$$

где a_i - фактическое, а y_i - предсказанное значение (здесь и далее).

Средний квадрат отклонения (MSE – Mean Squared Error)[5]:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |^2 .$$

Коэффициент детерминации (R^2)[4]:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} | a_i - y_i |^2}{\sum_{i=1}^{m} | \overline{y} - y_i |^2}, \qquad \overline{y} = \frac{1}{m} \sum_{i=1}^{m} y_i,$$

где $\overline{y}$ - среднее значение.

Получены следующие результаты.

Табл. 1

Значения метрик обученной модели

R^2	**0.036743**
MSE	**330.08631**
MAE	**2.6352**

Помимо посчитанных метрик было проведено сравнение результатов предсказания с результатами эвристически подобранных значений. В качестве подобранных значений были взяты результаты работы абстрактной модели, предсказывающей для любых входных сигналов одно и то же целое число. Данные модели далее будем называть константными и для простоты записи, далее будем обозначать модель, предсказывающую для любых входных сигналов число i, самим числом i. В табл. 2 показаны результаты оценки точности подобранных значений предсказаний.

Табл. 2

Сравнение значений метрик обученной и константных моделей

	0	**1**	**2**	**3**	**4**	**Модель**
MAE	3.0122	2.0122	2.2842	2.8951	3.6381	2.6352
MSE	51.751	346.726	343.702	342.677	343.653	330.086
R^2	-0.0264	-0.011	-0.0029	-4.4020	-0.0028	0.03674

По большинству метрик базовая модель превосходит эвристически подобранные предсказания. Значение метрик MAE и MSE для предсказаний нашей модели, в большинстве случаев, меньше, чем значение этих метрик для предсказаний константных моделей. Это означает, что средняя и среднеквадратичная ошибка предсказаний обученной модели меньше, чем у константных моделей, следовательно, созданная модель предсказывает точнее и работает лучше. Особенно это видно по метрике R^2, так как она меньше нуля для всех моделей, основанных на эвристически подобранные предсказаниях, а, значит, простое среднее будет давать результат лучше [4].

Табл.3

Анализ процентной ошибки обученной и константных моделей

	0	**1**	**2**	**3**	**4**	**Модель**
Ошибка в %	100 %	51 %	76 %	102 %	143 %	70 %

Дополнительно, для улучшения интерпретируемости результатов, был проведен сравнительный анализ предсказаний построенной модели и эвристически подобранных значений. Анализ проводился, как определение процентного соотношения между фактическими продажами маркетплейса и предсказаниями описанных выше моделей. Процентное соотношение рассчитывалось, как отношение прибыли маркетплейса в рублях и прибыли, рассчитанной на основе предсказаний моделей. Для удобства количество процентов прибыли, потерянное маркетплейсом из-за неверных предсказаний модели, будем называть процентной ошибкой. Результаты проведенного анализа показаны в табл. 3

Как видно из табл. 3, предложенная модель вполне эффективна и состоятельна. То, что константная модель, предсказывающая единицу для любых входных сигналов, имеет меньшую процентную ошибку, чем предложенная модель, объясняется спецификой полученного датасета – в исходном подмножестве товаров средние продажи были близки к единице.

Также была выделена важность признаков:

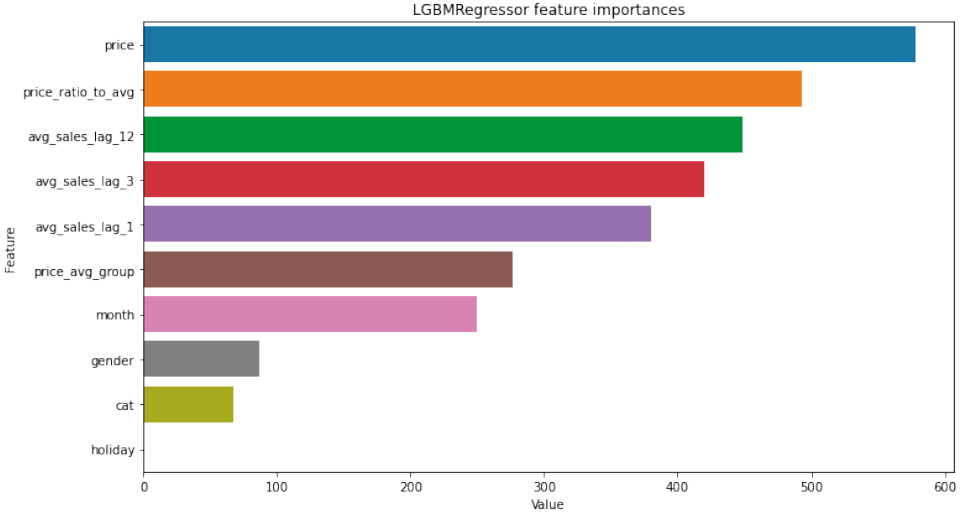

Рис. 4. Значения метрик обученной модели

На рис. 4 представлена диаграмма важности признаков для обученной модели. Данная диаграмма показывает, какими признаками руководствуется алгоритм при создании предсказания в большей или меньшей степени, соответственно. Иными словами диаграмма показывает то, насколько сильно выделенные нами признаки коррелируют с искомым значением. В рассматриваемом случае, анализируя диаграмму, можно убедиться, что наиболее значимыми признаками являются:

- price – цена товара;
- price_ratio_to_avg – отношение цены товара, к средней цене товара в группе;
- avg_sales_lag_12 – средние продажи в группе, к которой принадлежит товар, год назад;

- avg_sales_lag_3 – средние продажи в группе, к которой принадлежит товар, 3 месяца назад;
- avg_sales_lag_1 – средние продажи в группе, к которой принадлежит товар, месяц назад;
- price_avg_group – средняя цена в группе, к которой принадлежит товар;
- month – месяц, в котором товар вышел на торговую площадку;
- cat – категория, к которой принадлежит товар.

4. ВЫВОД

На данном этапе развития работы было представлено два результата: вычислен примерный убыток интернет-маркетплейса «Озон», а также построена базовая модель, обученная на данных, представленными «Озоном». Данная модель будет основой для дальнейших исследований задачи предсказания спроса на товары, не имеющие истории продаж, и будет использоваться в ансамбле вместе с другими моделями, реализующими другие подходы к решению данной задачи.

Литература

1. *Bisong E.* Introduction to Scikit-learn // Building Machine Learning and Deep Learning Models on Google Cloud Platform 2019. P. 215–229.
2. *Cerda P., Varoquaux G., Kégl B.* Similarity encoding for learning with dirty categorical variables // Machine Learning. 2018. P. 1477–1494.
3. *Ke G.* et al. Lightgbm: A highly efficient gradient boosting decision tree // Advances in neural information processing systems. 2017. P. 3146–3154.
4. *Redell N.* Shapley Decomposition of R-Squared in Machine Learning Models // arXiv preprint arXiv:1908.09718. 2019.
5. *Botchkarev, Alexei.* “Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology.” // arXiv preprint arXiv:1809.03006. 2018.
6. *Al Daoud E.* Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset // International Journal of Computer and Information Engineering. 2019. P. 6–10.

Using Machine Learning Methods to Solve Problems of Forecasting Demand for New Products in the Internet Marketplace

Artem A. Osin*
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0002-2664-1370
e-mail: artemosin1@yandex.ru

Artem K. Fomin**
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0003-3545-4435
e-mail: artem.fomin@outlook.com

Gleb B. Sologub***
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

Vladimir I. Vinogradov****
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0003-3773-9653
e-mail: vvinogradov@inbox.ru

The work is aimed at researching the possibility of using machine learning methods to build models for forecasting demand for new products in the online store Ozon.ru. Approaches to the solution that were not previously used in a specific task are proposed for consideration. Data on sales history and storage of goods at Ozon.ru are used as a sample. There is a description and analysis of the approximate loss of

For citation:
Osin A.A., Fomin A.K., Sologub G.B., Vinogradov V.I. Using Machine Learning Methods to Solve Problems of Forecasting Demand for New Products in the Internet Marketplace. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2020. Vol. 10, no. 4, pp. 41–50. DOI: https://doi.org/10.17759/mda.2020100404 (In Russ., abstr. in Engl.).

***Artem A. Osin,** Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0002-2664-1370, e-mail: artemosin1@yandex.ru

****Artem K. Fomin,** Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0003-3545-4435, e-mail: artem.fomin@outlook.com

*****Gleb B. Sologub,** PhD (Physics and Mathematics), Associate Professor, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

******Vladimir I. Vinogradov,** PhD (Physics and Mathematics), Associate Professor, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0003-3773-9653, e-mail: vvinogradov@inbox.ru

the Ozon.ru website, the data used, the process of building a base model, and the results obtained. It describes the metrics used to evaluate the prediction results and makes a comparative analysis between the prediction results of the built model and the results of heuristically selected values.

Keywords: demand forecasting, new products, encoding, gradient busting, regression, preprocessing, data processing, machine learning.

References

1. Bisong E. Introduction to Scikit-learn // Building Machine Learning and Deep Learning Models on Google Cloud Platform 2019. P. 215–229.
2. Cerda P., Varoquaux G., Kégl B. Similarity encoding for learning with dirty categorical variables // Machine Learning. 2018. P. 1477–1494.
3. Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree // Advances in neural information processing systems. 2017. P. 3146–3154.
4. Redell N. Shapley Decomposition of R-Squared in Machine Learning Models // arXiv preprint arXiv:1908.09718. 2019.
5. Botchkarev, Alexei. “Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology.” // arXiv preprint arXiv:1809.03006. 2018.
6. Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset // International Journal of Computer and Information Engineering. 2019. P. 6–10.