ПРОГНОЗИРОВАНИЕ ПРОДАЖ В РИТЕЙЛЕ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Е. А. Артемьев, email: artemev.15.01.1998@list.ru, В.В. Мокшин

Казанский национальный исследовательский технический университет

Аннотация. *В современном мире ритейла недостаточно завозить продукцию в торговые точки, выставлять её на стеллажи. Отчётности от мерчендайзеров или по продажам за месяц стали хорошим инструментом анализа, но это требует времени и больших человеческих ресурсов. Эффективнее использовать машинное обучение для прогнозирования продаж на месяцы вперёд, позволяя заранее принимать решения по кол-ву, типам и ценам товаров в торговых сетях. В данной работе вы узнаете, как реализовать модель машинного обучения, которая может прогнозировать продажи в зависимости от определенных характеристик, таких как день продажи, торговая точка, стоимость, кол-во проданных товаров за предыдущие дни и категория товара.*

Ключевые слова: *временное прогнозирование, регрессионный анализ, корреляционный анализ, регрессия, продажи, ритейл.*

Введение

Ритейл – это механизм продажи товаров и услуг, запутанная и сложная сеть из покупателей и продавцов. Продавцы закупают товары в определённом кол-ве оптом и продают по определённой цене на различных площадках и в различных розничных сетях.

Покупатели берут выложенный товар исходя из своих мотивов. Собрав датасет из проданных товаров, можно понять мотивы людей, месяцы высокого спроса и малого, интересные продукты и не особо в определённые периоды.

Наша цель в данной статье — изучить прогнозирование будущих продаж основе прошлых продаж. Эти данные поступают с веб-сайта Kaggle, на котором собраны датасеты различных компаний. Важно иметь огромный массив данных для того, чтобы модель могла вычленить зависимости между данными.

1. Методология

Данные были получены с веб-сайта Kaggle; набор данных был озаглавлен “Predict Future Sales”. Перед началом работы с данными

© Артемьев Е.А., Мокшин В.В.

были выдвинуты несколько различных гипотез относительно того, какие переменные будут иметь большое значение для увеличения продаж. Наиболее интересными для обучения признаками были выделены цена, дата, торговая точка, сам товар. После первых экспериментов стало ясно, что этих параметров недостаточно и их использование в том виде, в котором они представлены в датасете неудачно и требует преобразования.

Данные были разделены на обучающуюся и тестовую выборки в отношении 80:20 и очищены путём удаления дубликатов, выбросов, строк с пустыми значениями, сортировки по дате, замены дат и строковых значений числовыми для более эффективного обучения. Затем был проведён корреляционный анализ, чтобы определить ключевые факторы, которые имеют тесную связь с предсказанием поля, отвечающего за кол-во проданных товаров в определённый промежуток времени. Связь между данными была установлена крайне незначительной, для чего было принято перегруппировать данные, но после проведения экспериментов с линейной регрессией. Алгоритм линейной регрессии сработал неудачно, но дал нам важную информацию насчёт влияния параметров на обучение: id товара, id торговой точки и месяц продаж. Затем заменили алгоритм на метод ансамблевого обучения Extreme Gradient Boosting Regressor для адаптивного изменения распределения обучающих данных, уделяя больше внимания ранее неправильно классифицированным записям для создания базовых учащихся. Алгоритм показал себя лучше, указав нам на то, что всё те же поля оказывают наибольшее влияние на обучение и на то, что можно перегруппировать данные и добавить новые признаки.

Для оценки точности алгоритма были использованы такие метрики как MAE, MSE, RMSE. MAE – В статистике, средняя абсолютная ошибка является мерой ошибок между парными наблюдениями, выражающих то же явление. Вычисляется по формуле:

$$MAE = \sum_{i=1}^{n} \left| y_i - x_i \right| / n = \sum_{i=1}^{n} \left| e_i \right| / n ,$$

– n – число прогнозов;

– y – спрогнозированное значение;

– x – изначальные значения.

Средняя абсолютная ошибка известна как мера точности, зависящая от масштаба и имеет вид:

$$MSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 ,$$

– n – число прогнозов;
– y – изначальные данные;
– $\hat{y}$ - предсказанное значение.

RMSE (Root Mean Square Error) - это квадратный корень из значения, полученного с помощью функции среднеквадратической ошибки.

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\Pr edicted_i - Actual_i \right)^2 / N} ,$$

– $\Pr edicted$ – число прогнозов;
– $Actual$ – изначальные данные;
– N - предсказанное значение.

2. Анализ

Поскольку мы пробуем предсказать кол-во проданных товаров в определённый промежуток времени, мы использовали регрессионный анализ. Регрессионный анализ — набор статистических методов исследования влияния одной или нескольких независимых переменных на зависимую.

В исследовании были некоторые ограничения. Одним из ограничений являются выбросы и определение переменных, необходимых для обучения. Для выявления второго был использован корреляционный анализ. По полученным данным видно, что связь крайне несущественная, поэтому были выбраны все признаки для первых экспериментов (рис. 1).

	date_block_num	shop_id	item_id	item_price	item_cnt_day
date_block_num	1.000000	0.019273	0.009356	0.095010	0.009402
shop_id	0.019273	1.000000	0.029396	-0.024034	-0.005230
item_id	0.009356	0.029396	1.000000	-0.134104	0.016650
item_price	0.095010	-0.024034	-0.134104	1.000000	0.011197
item_cnt_day	0.009402	-0.005230	0.016650	0.011197	1.000000

Рис. 1. Корреляционный анализ

После пропуска датасета по линейной регрессии и методу ансамблевого обучения стало ясно, что сам товар и торговая точка влияют на обучение в большей степени (рис. 2). Но результаты обучения были не оптимистичными, поэтому было принято решение сгруппировать данные по-другому и связь между признаками стала выше (рис. 3) [1].

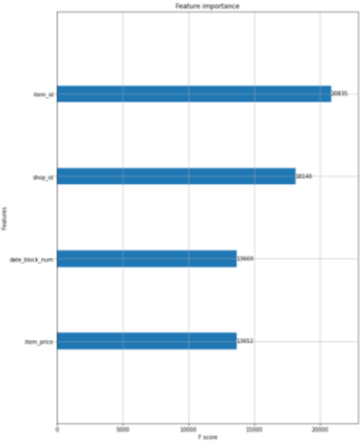

Рис. 2. Параметры, влияющие на обучение

index	date_block_num	shop_id	item_id	item_cnt_month	shop_category	shop_city	item_category_id
date_block_num	1.0	0.022619083075971 48	-0.012621322296694718	-0.006062333539953814	-0.03221215477246299	0.021269480301459403	0.016756255619110706
shop_id	0.022619083075971 48	1.0	-0.0003466411615666085	0.008252940600719245	0.10051819869574335	0.9861300305824917	-0.0001956849529621953 5
item_id	-0.012621322296694718	-0.0003466411615666085	1.0	-0.012189604917996203	0.0003326463236246044 6	-0.0002903594524420263 6	0.12937494419269333
item_cnt_month	-0.006062333539953814	0.008252940600719245	-0.012189604917996203	1.0	-0.0016701449339382713	0.007829367447621534	-0.026884647916516554
shop_category	-0.03221215477246299	0.10051819869574335	0.0003326463236246044 6	-0.0016701449339382713	1.0	0.05468221124064191	-0.0009358160273374369
shop_city	0.021269480301459403	0.9861300305824917	-0.0002903594524420263	0.007829367447621534	0.05468221124064191	1.0	-0.0001388460216307458
item_category_id	0.016756255619110706	-0.0001956849529619535	0.12937494419269333	-0.025884647916516554	-0.0009358160273374369	-0.0001388460216307458	1.0
name2	0.020127533224157122	0.008056889732276982	-0.24640304573159538	0.027711857818949436	-0.0005264722884472049	0.000754746281642803 6	-0.2918051222911194
name3	-0.07904698988103075	-0.0016255289111336167	0.11032922760126965	-0.017796056904993 73	0.002839935292979115	-0.0015198376723119679	-0.14550194355133086
subtype_code	0.10084423034516954	0.002112556427581241 6	-0.13234324815627913	0.025074767496777888	-0.0034724262146504114	0.001940215826414798	0.26415979556212584
type_code	0.004441461634530 57	-0.0004401424940894664	0.14784008099308075	-0.036463981241664445	-0.0016206289064894375 8	-0.0003994706486555426	0.9564247277181361
item_cnt_month_lag_1	0.008136065869174807 2	0.010093317723671287	-0.015997869018976072	0.7267360094107479	3.134869149432866e-05	0.009469280436412323	-0.029674942216121177
item_cnt_month_lag_2	2.6156119589219488e-05	0.010809772368098129	-0.017458546536828894	0.6496957412041661	0.001681233239168735	0.010117837634381502	-0.031909922969952825
item_cnt_month_lag_3	0.008125657857800271	0.011386836136480571	-0.018325438094583365	0.6268790840764353	0.001632588774289238	0.010695152398431555	-0.033169386422163211
date_avg_item_cnt_lag_1	-0.06726472062110475	-0.0004663405563231224	-0.004123706168004966	0.008700278330489985	0.002858366364921 88	-0.0002739059766749846	0.009171835566500112
date_item_avg_item_cnt_lag_1	4.741952703157681 5e-05	0.5304731560253376e-05	-0.024923207327468548	0.5181497817247801	4.232229953194888e-05	0.024154831207095e-05	-0.04635048308840784
date_item_avg_item_cnt_lag_2	-0.000175508625804279 27	2.160494758278723e-05	-0.026808623394318855	0.4698146499633793	0.000152987095202049 82	3.210160527709949e-05	-0.048767395349483184
date_item_avg_item_cnt_lag_3	0.008203186022966172 7	9.836313673608655e-05	-0.028142740768615154	0.462792129367675	8.005466146288325e-05	0.000112467399858937 35	-0.050642634081167 13
date_shop_avg_item_cnt_lag_1	-0.012450868638496523	0.122096107388192 43	-0.000873466513361645 2	0.07071063884555667	-0.014627520275031634	0.11570357992610378	0.000320879921499432 6
date_shop_avg_item_cnt_lag_2	-0.016934454504594944	0.127048421269426 03	9.381293799732618e-05	0.060774337673498 66	-0.008168556810266568	0.120409960970261 52	0.000781657016784 3007
date_shop_avg_item_cnt_lag_3	-0.022102500381443537	0.131424233028721 3	0.000592849183154542 7	0.058247863918686 77	0.001990752391113 17	0.124674569422139 1	0.000561913952285 1313
date_shop_item_avg_item_cnt_lag_1	0.008136065869174807 2	0.010093317723671287	-0.015997869018976072	0.7267360094107479	3.134869149432866e-05	0.009469280436412323	-0.029674942216121177
date_shop_item_avg_item_cnt_lag_2	2.6156119589219488e-05	0.010809772368098129	-0.017458546536828894	0.6496957412041661	0.001681233239168735	0.010117837634381502	-0.031909922969952825
date_shop_item_avg_item_cnt_lag_3	0.008125657857800271	0.011386836136480571	-0.018325438094583365	0.6268790840764353	0.001632588774289238	0.010695152398431555	-0.033169386422163211
date_shop_subtype_avg_item_cnt_lag_1	0.009787731172583698	0.010466493929225 59	-0.014067686057534 44	0.7329280427032616	6.599841884076423e-07	0.009736751307852454	-0.033021299332201277
date_city_avg_item_cnt_lag_1	-0.019002590573999965	0.139248863421236 38	-0.001185274669225323	0.04928373782776141	-0.13114500778137728	0.163878907162400 17	0.000277807386447162 77
date_item_city_avg_item_cnt_lag_1	1.017500162092536 7e-05	0.009325469899054153	-0.010038009928411 72	0.6395049237461435	-0.008099519918872072	0.011004113758666665	-0.033636188982257 97
delta_price_lag	-0.146137338186137 56	-0.009523059971715654	-0.007435196042829805	0.006677799799355643	0.003727293151094179	-0.005055206394239604	0.047132511587313 72
delta_revenue_lag_1	-0.013851158866880 07	0.050878576943795426	-0.001810379340446 3623	0.061037838527679 14	-0.027706801057245 8	0.041994376299156 9	0.000829236938418 717
month	0.023529407680652724	-0.006616582211597 59	-0.003335899796496866 5	0.009486422419597 05	-0.009893586160722648	-0.005994904818542 95	0.014603361203244956
days	-0.059713427999888 04	-0.003056837086968433	0.003705210543714814	0.028121508544980199	-0.005694150297368665	-0.002948135447137487	-5.128907334656071e-05
item_shop_first_sale	0.461056050291655 6	-0.018094646344466689	-0.008633745669777614	0.027630936009862946	0.009905368862980527	-0.015330427923586047	-0.003819419978683615
item_first_sale	0.481935065829028 45	0.012278577488492481	-0.009334610057696969	-0.032419860870183954	-0.014640328534424303	0.011560095079999759	-0.003416577162489271

Рис. 3. Корреляционный анализ преобразованного датасета

3. Прогнозирование

Мы пробуем предсказать кол-во проданных товаров в определённый промежуток времени. Для этого мы используем несколько алгоритмов обучения модели: «Линейная регрессия», “eXtreme Gradient Boosting Regressor”.

Применив линейную регрессию, получаем неутешительные результаты. Мы видим такие метрики как MAE, MSE, RMSE и таблицу, по которой можем увидеть, какие столбцы в какой степени влияют на обучение (рис. 4).

```
mse: 18.373, mae: 0.523
Mean Absolute Error: 0.5229426436550997
Mean Squared Error: 18.3727463847258
Root Mean Squared Error: 4.286344174786458
```

	Coefficient
date_block_num	0.004828
shop_id	-0.000826
item_id	0.000008
item_price	0.000021

Рис. 4. Результаты линейной регрессии

Визуализировав данные для ТТ с id=2 в первый месяц 2015 года, видим, насколько плохо справилась модель, если смотреть в разрезе ТТ и периода (рис. 5). Дать советы магазину по такому результату не получится.

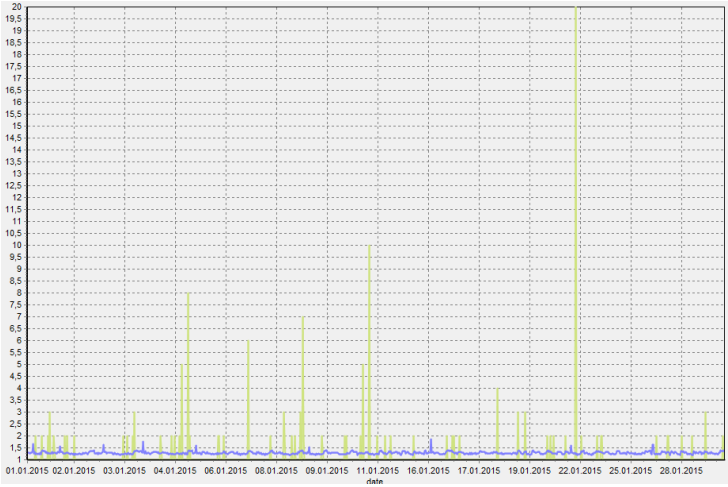

Рис. 5. График полученных и исходных данных

EXtreme Gradient Boosting Regressor - это метод ансамблевого обучения, который позволит создать окончательную модель путем объединения нескольких слабых моделей. Как описано в названии машины для повышения градиента, используется метод градиентного спуска и повышения. Метод Boosting использует итеративную процедуру для адаптивного изменения распределения обучающих данных, уделяя больше внимания ранее неправильно классифицированным записям для создания базовых учащихся. Это можно рассматривать как последовательное добавление моделей до тех пор, пока не прекратятся дальнейшие улучшения. Для минимизации потерь используется алгоритм градиентного спуска. При повышении градиента, когда прогнозы нескольких моделей комбинируются, градиент используется для оптимизации прогнозирования усиленной модели в каждом раунде повышения.

Мы также взяли отсортированный по дате датасет и разделили его на обучающую и тестовую выборки. И подали его на вход модели без ухищрений. Получаем результаты, которые выглядят лучше, чем у линейной регрессии, но недостаточно хороши, чтобы использовать в реальной жизни.

Результаты обучения выглядят следующим образом (рис. 6):

– [1] validation_0-rmse:2.09102 validation_1-rmse:2.09102

– [2] validation_0-rmse:2.05408 validation_1-rmse:2.05408
– [3] validation_0-rmse:2.03131 validation_1-rmse:2.03131
– [4] validation_0-rmse:2.00578 validation_1-rmse:2.00578
– [5] validation_0-rmse:1.98453 validation_1-rmse:1.98453
– [6] validation_0-rmse:1.96968 validation_1-rmse:1.96968
– …
– [998] validation_0-rmse:1.70407 validation_1-rmse:1.70407
– [999] validation_0-rmse:1.70392 validation_1-rmse:1.70392

Mean Absolute Error: 0.45253064959522105
Mean Squared Error: 16.83557398677999
Root Mean Squared Error: 4.103117593584175

	Actual	Predicted
2348679	1.0	1.464342
2348680	2.0	1.464342
2348681	1.0	1.464342
2348682	1.0	1.464342
2348683	1.0	2.632516
...	...	...
2935844	1.0	1.029175
2935845	1.0	1.029175
2935846	1.0	1.107367
2935847	1.0	1.029175
2935848	1.0	1.029175

587170 rows × 2 columns

Рис. 6. Результаты EXtreme Gradient Boosting Regressor

Feature_importances позволяет оценить, какие признаки оказывали существенное влияние на обучение (рис. 7).

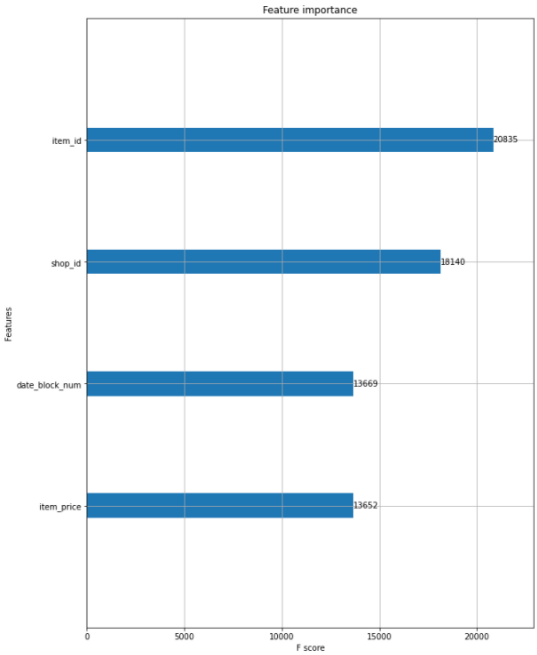

Рис. 7. Feature_importances. EXtreme Gradient Boosting Regressor

Попробуем уменьшить ошибку, доработав алгоритм выше путём группировки данных и добавлением новых признаков. Результаты обучения (рис. 8):

– [0] validation_0-rmse:1.18496 validation_1-rmse:1.11945
– [1] validation_0-rmse:1.12885 validation_1-rmse:1.07960
– [2] validation_0-rmse:1.08653 validation_1-rmse:1.04936
– [3] validation_0-rmse:1.04426 validation_1-rmse:1.02158
– [4] validation_0-rmse:1.01560 validation_1-rmse:0.99966
– [5] validation_0-rmse:0.98959 validation_1-rmse:0.98105
– …

– [67] validation_0-rmse:0.76339 validation_1-rmse:0.91226
– [68] validation_0-rmse:0.76288 validation_1-rmse:0.91230
– [69] validation_0-rmse:0.76230 validation_1-rmse:0.91229
– [70] validation_0-rmse:0.76176 validation_1-rmse:0.91222

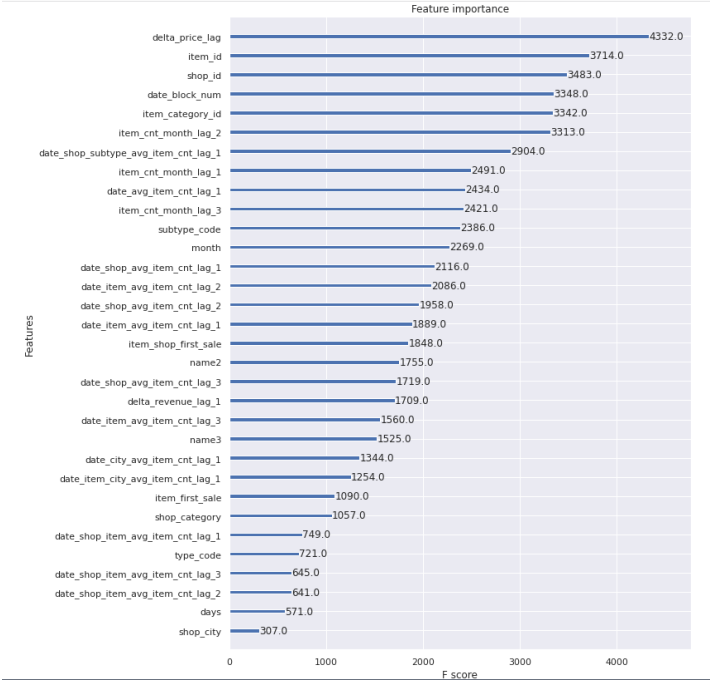

Рис. 8. Feature_importances. EXtreme Gradient Boosting Regressor

Посмотрим, насколько хорошо предсказала модель кол-во проданных товаров за январь 2015 года в ТТ с id=50.

По диаграмме (рис. 9) видно, что результаты несильно отличаются, но опять же стоит отметить, что результаты далеки от совершенных и предсказывать по такой модели опасно.

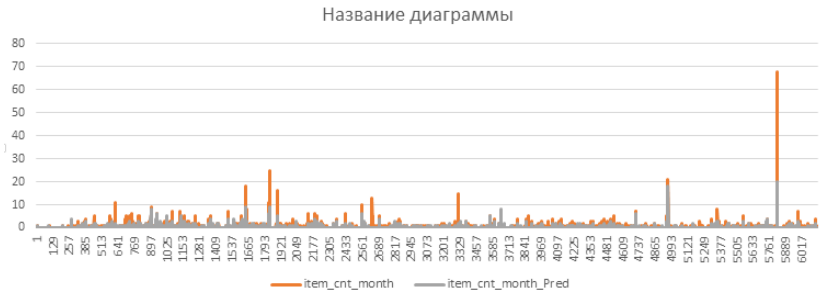

Рис. 9. График исходных и полученных данных EXtreme Gradient Boosting Regressor

Заключение

В данном исследовании представлен подход, основанный на машинном обучении, для оценки продаж на будущие месяцы по данным прошлых месяцев. Полученная модель требует улучшений и, возможно, замены на более сложные модели с обширным функционалом, но положительный результат использования машинного обучения в ритейле виден из графиков выше.

Большой набор данных был обработан, прежде чем передать его предлагаемым классификаторам и спрогнозировать кол-во проданных товаров. Необходимо для каждой задачи правильно сгруппировать данные и выделить необходимые признаки. Оценка основана на сравнении результатов предсказания и реальных значений продаж, производительности модели были оценены с помощью метрик MAE, MSE, RMSE, что показало, насколько модель приближена к реальной. По полученным данным, можно сказать, что разработанный алгоритм имеет место быть для грубого предсказания продаж как подспорье для дальнейшего анализа и выводов. Также результаты указывают на то, что анализ данных по продажам – это непростая задача с множеством подводных камней.

Литература

1. Мокшин, В. В. Рекурсивный алгоритм построения регрессионных моделей сложных вероятностных объектов / В. В. Мокшин, И. Р. Сайфудинов, А. П. Кирпичников // Вестник Технологического университета. – 2017. – № 9. – С. 112-116.

2. Мокшин, В. В. Метод формирования модели анализа сложной системы / В. В. Мокшин, И. М. Якимов // Информационные технологии. – 2011. – № 5. – С. 46-51.