**«Цены на туры»**

**Описание задачи**:

Вам необходимо разработать прогноз продаж для туроператора, продающего туры в Объединенные Арабские Эмираты с помощью статистических методов.

**Дано**:

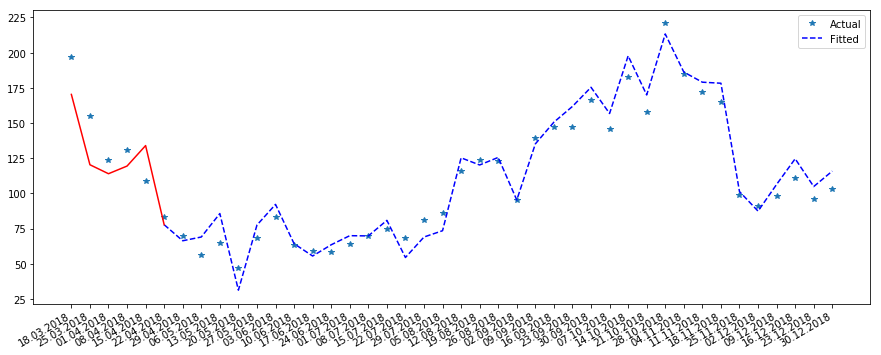
1. Туроператор «ОАЭ Тур» продает туры в ОАЭ на своем сайте;
2. Единственным каналом продвижения является SEO в Яндексе. Единственным поисковым запросом, по которому оптимизирован сайт, является «туры в оаэ»;
3. У «ОАЭ Тур» есть только один конкурент;
4. Основной фактор конкуренции – это цена;
5. Имеется статистика по средним ценам ну тур у «ОАЭ Тур» и конкурента, а также количеству заказов «ОАЭ Тур», в недельном разрезе.

**Найти**:

1. Определить ключевой рыночный фактор, позволяющий релевантно оценить рыночный спрос на туры в ОАЭ в целом по рынку;
2. Найти значение этого фактора понедельно на весь период наблюдений;
3. Сформировать модель, описывающую зависимость между спросом, ценами и количеством заказов. Привести обоснование модели;
4. С помощью нее рассчитать понедельный прогноз продаж на тестовый период.

**Решение**:

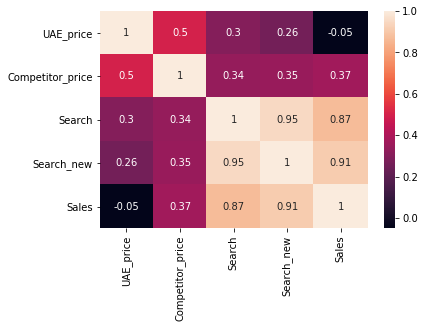
Необходимо улучшить качество прогноза. Построенная модель плохо описывает изменчивость исходного ряда в начале года – прогноз необходимо рассчитать именно на этом участке. Постараемся это исправить и улучшить качество модели:



Вернемся к внешнему фактору, использованному для оценки рыночного спроса на туры – статистике Яндекса по запросу «туры в ОАЭ». Попробуем улучшить информационную значимость этого фактора – соберем статистику по запросу «туры в оаэ», то есть будем учитывать показы, не включающие предлог «в». Обновленная статистика выглядит следующим образом:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Период продаж** | **История показов** | **Период продаж** | **История показов** | **Период продаж** | **История показов** |
| 12.03.2018 - 18.03.2018 | 12010 | 18.06.2018 - 24.06.2018 | 3907 | 24.09.2018 - 30.09.2018 | 11274 |
| 19.03.2018 - 25.03.2018 | 10692 | 25.06.2018 - 01.07.2018 | 4165 | 01.10.2018 - 07.10.2018 | 12418 |
| 26.03.2018 - 01.04.2018 | 10240 | 02.07.2018 - 08.07.2018 | 5356 | 08.10.2018 - 14.10.2018 | 12044 |
| 02.04.2018 - 08.04.2018 | 6661 | 09.07.2018 - 15.07.2018 | 5644 | 15.10.2018 - 21.10.2018 | 11432 |
| 09.04.2018 - 15.04.2018 | 5799 | 16.07.2018 - 22.07.2018 | 5844 | 22.10.2018 - 28.10.2018 | 11197 |
| 16.04.2018 - 22.04.2018 | 4785 | 23.07.2018 - 29.07.2018 | 6289 | 29.10.2018 - 04.11.2018 | 10776 |
| 23.04.2018 - 29.04.2018 | 3769 | 30.07.2018 - 05.08.2018 | 6865 | 05.11.2018 - 11.11.2018 | 10346 |
| 30.04.2018 - 06.05.2018 | 3509 | 06.08.2018 - 12.08.2018 | 7583 | 12.11.2018 - 18.11.2018 | 10107 |
| 07.05.2018 - 13.05.2018 | 3528 | 13.08.2018 - 19.08.2018 | 7816 | 19.11.2018 - 25.11.2018 | 8593 |
| 14.05.2018 - 20.05.2018 | 3688 | 20.08.2018 - 26.08.2018 | 7742 | 26.11.2018 - 02.12.2018 | 8192 |
| 21.05.2018 - 27.05.2018 | 3646 | 27.08.2018 - 02.09.2018 | 7311 | 03.12.2018 - 09.12.2018 | 8618 |
| 28.05.2018 - 03.06.2018 | 3951 | 03.09.2018 - 09.09.2018 | 8307 | 10.12.2018 - 16.12.2018 | 7824 |
| 04.06.2018 - 10.06.2018 | 3949 | 10.09.2018 - 16.09.2018 | 9133 | 17.12.2018 - 23.12.2018 | 7204 |
| 11.06.2018 - 17.06.2018 | 3961 | 17.09.2018 - 23.09.2018 | 10801 | 24.12.2018 - 30.12.2018 | 6039 |
|  |  |  |  |  |  |
| **Период продаж** | **История показов** |
| 31.12.2018 - 06.01.2019 | 6924 |
| 07.01.2019 - 13.01.2019 | 10277 |
| 14.01.2019 - 20.01.2019 | 10666 |
| 21.01.2019 - 27.01.2019 | 9304 |
| 28.01.2019 - 03.02.2019 | 8975 |
| 04.02.2019 - 10.02.2019 | 9349 |
| 11.02.2019 - 17.02.2019 | 8600 |
| 18.02.2019 - 24.02.2019 | 8540 |
| 25.02.2019 - 03.03.2019 | 9186 |

Посмотрим на корреляционную матрицу с учетом нового фактора (название фактора в матрице – “Search\_new”):



Новая статистика по запросам лучше коррелирует с целевой переменной (r = 0,91). При этом, корреляция с остальными факторами небольшая. Включим новый фактор в модель линейной регрессии и посмотрим на результаты (ранее использованный фактор “Search” исключен):

|  |  |  |  |
| --- | --- | --- | --- |
| constant | b1 | b2 | b3 |
| -21.208 | -0.00102 | 0.00128 | 0.0143 |

Качество на тренировочной выборке:

The model performance for training set

--------------------------------------

RMSE is 9.404753947719632

R2 score is 0.9556066464849917

Качество на тестовой выборке:

The model performance for testing set

--------------------------------------

RMSE is 8.798975058102714

R2 score is 0.9548183631086457

Качество на кросс-валидации:

Cross\_val\_score

--------------------------------------

Train cross\_val\_score R2

[0.9874645 0.968065 0.79649955 0.96202887 0.95742486] 0.934297

Test cross\_val\_score R2 [

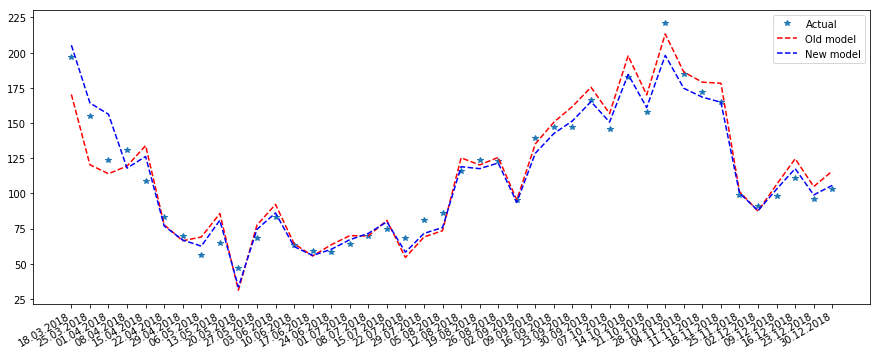
0.94189403 0.96181692 0.9643125 0.96913625 0.85140723] 0.9377133

Статистика модели:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Sales | **R-squared:** | 0.956 |
| **Model:** | OLS | **Adj. R-squared:** | 0.950 |
| **Method:** | Least Squares | **F-statistic:** | 172.2 |
| **Date:** | Sun, 12 May 2019 | **Prob (F-statistic):** | 2.31e-16 |
| **Time:** | 20:17:04 | **Log-Likelihood:** | -102.48 |
| **No. Observations:** | 28 | **AIC:** | 213.0 |
| **Df Residuals:** | 24 | **BIC:** | 218.3 |
| **Df Model:** | 3 |  | |
| **Covariance Type:** | nonrobust |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | -21.208 | 18.540 | -1.144 | 0.264 | -59.473 | 17.056 |
| **UAE\_price** | -0.0010 | 0.000 | -9.221 | 0.000 | -0.001 | -0.001 |
| **Competitor\_price** | 0.0013 | 0.000 | 5.569 | 0.000 | 0.001 | 0.002 |
| **Search\_New** | 0.0143 | 0.001 | 17.249 | 0.000 | 0.013 | 0.016 |

Новая модель показала лучшие значения R2, RMSE на тренировочной, тестовой выборках и в кросс-валидации. Значения p-value при коэффициентах также позволяют отвергнуть гипотезу о незначимости. Модель заметно лучше описывает изменчивость исходного ряда, особенно в начале года:



Попытаемся добиться лучшего качества с помощью более сложных алгоритмов. Построим модель гребневой регрессии. В качестве матрицы X будем использовать матрицу полиномиальных признаков второй степени.

Для лучшей генерации матрицы полиномиальных признаков преобразуем факторы с сильной корреляцией между друг другом: «Средняя цена «ОАЭ ТУР», «Средняя цена конкурента».

Посмотрим на корреляционную матрицу преобразованных признаков: разности, средней, отношения и логарифма суммы двух признаков. Наиболее информативным признаком является разность - ее и будем использовать для генерации полиномиальных признаков:



Наилучшие результаты были получены с матрицей, сгенерированной с параметром interaction\_only = **True,** то есть для создания матрицы не были использованы признаки вида X12, X22 и т.д. В нашем случае матрица выглядит следующим образом:

|  |  |  |  |
| --- | --- | --- | --- |
| **One** | **X1** | **X2** | **X1\*X2** |
| 1.0 | -3321.15 | 4165.0 | -13832589 |
| 1.0 | -14247.25 | 10776.0 | -153528366 |
| 1.0 | 9489.40 | 5356.0 | 50825226 |
| 1.0 | 30486.50 | 8593.0 | 261970494 |
| 1.0 | -7854.90 | 10107.0 | - 79389474 |

Также сделаем нормировку признаков, чтобы учесть масштаб их изменения и сделать штраф регуляризатора более честным. Рассчитаем коэффициенты модели гребневой регрессии:

*(alpha = 0.4)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| constant | b1 | b2 | b3 | b4 |
| -2.89e-17 | 0.0 | -0.10448035 | 0.81969735 | -0.30263959 |

Из-за номировки коэффициент b1 и константа получились равными нулю.

Качество на тренировочной выборке:

The model performance for training set

--------------------------------------

RMSE is 0.20500772275020612

R2 score is 0.9579718336127746

Качество на тестовой выборке:

The model performance for testing set

--------------------------------------

RMSE is 0.11455679714074682

R2 score is 0.9847413386063385

Качество на кросс-валидации:

Cross\_val\_score

--------------------------------------

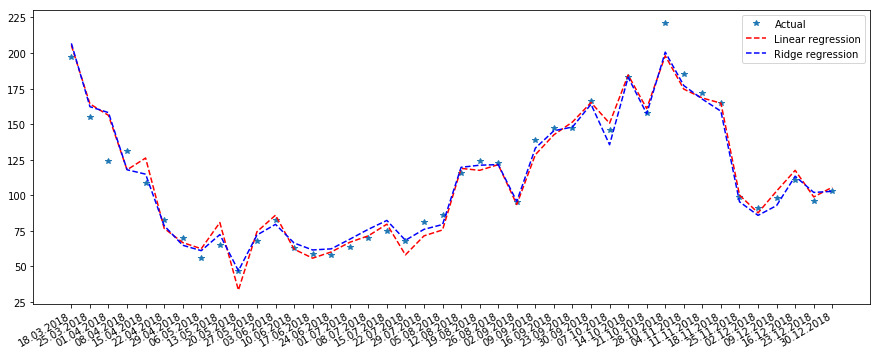
Train cross\_val\_score R2

[0.98085806 0.97406972 0.81549945 0.99194395 0.97191127] 0.946857

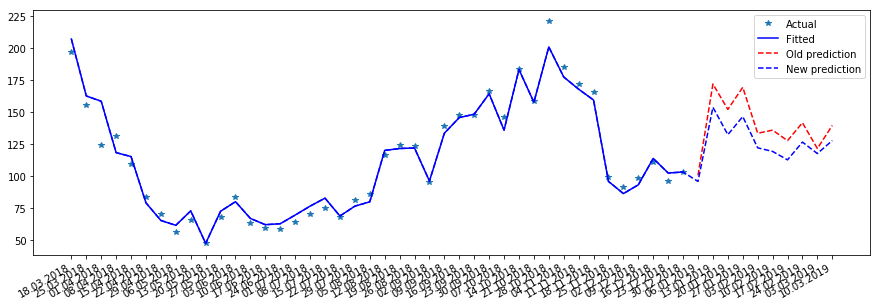
Test cross\_val\_score R2

[0.96413971 0.97032098 0.98609787 0.98299475 0.75868684] 0.932448

Новая модель показала неплохие результаты. Качество на тестовой выборке получилось выше, чем на тренировочной – это может быть вызвано особенностью разбиения, а также механизмом регуляризации: модель не переобучается на тренировочной выборке и из-за этого ошибка получается чуть больше, при этом на более маленькой тестовой выборке модель срабатывает немного лучше. Коэффициенты детерминации на кросс-валидации также получились выше.



С помощью модели рассчитаем понедельный прогноз продаж на тестовый период:



|  |  |  |  |
| --- | --- | --- | --- |
| Период продаж | Средняя цена «ОАЭ ТУР», руб./заказ | Средняя цена конкурента, руб./заказ | Прогноз продаж, заказов |
| 31.12.2018 - 06.01.2019 | 106651,7 | 96013,79 | 95 |
| 07.01.2019 - 13.01.2019 | 101673,5 | 102984,4 | 153 |
| 14.01.2019 - 20.01.2019 | 123099,7 | 103358,1 | 132 |
| 21.01.2019 - 27.01.2019 | 119102,7 | 125948,5 | 146 |
| 28.01.2019 - 03.02.2019 | 101896,1 | 91143,04 | 122 |
| 04.02.2019 - 10.02.2019 | 124413,5 | 107045,3 | 119 |
| 11.02.2019 - 17.02.2019 | 125538,7 | 110561,7 | 112 |
| 18.02.2019 - 24.02.2019 | 111693,3 | 110301,8 | 126 |
| 25.02.2019 - 03.03.2019 | 112086,9 | 95005,36 | 117 |
| 04.03.2019 - 10.03.2019 | 85283,17 | 103556 | 127 |