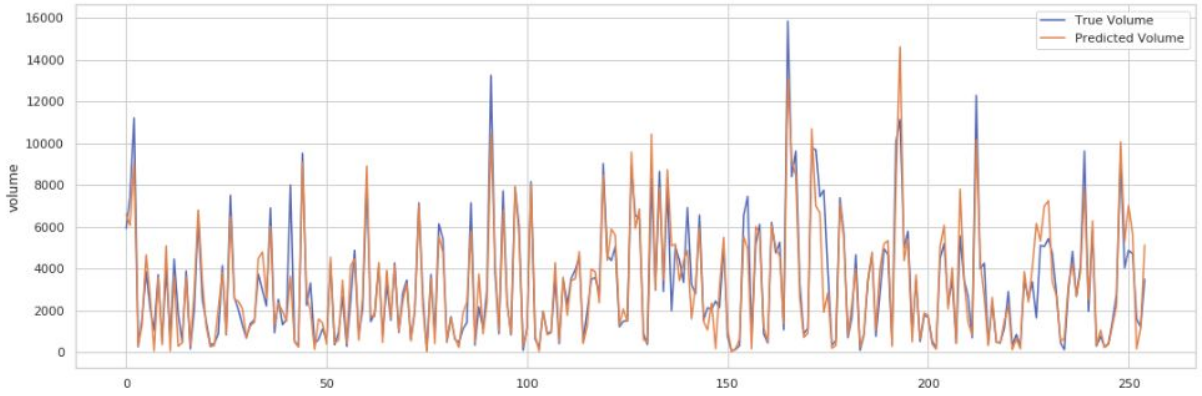
Краткий отчет для GPN Intelligence CUP направление Продвинутая Аналитика

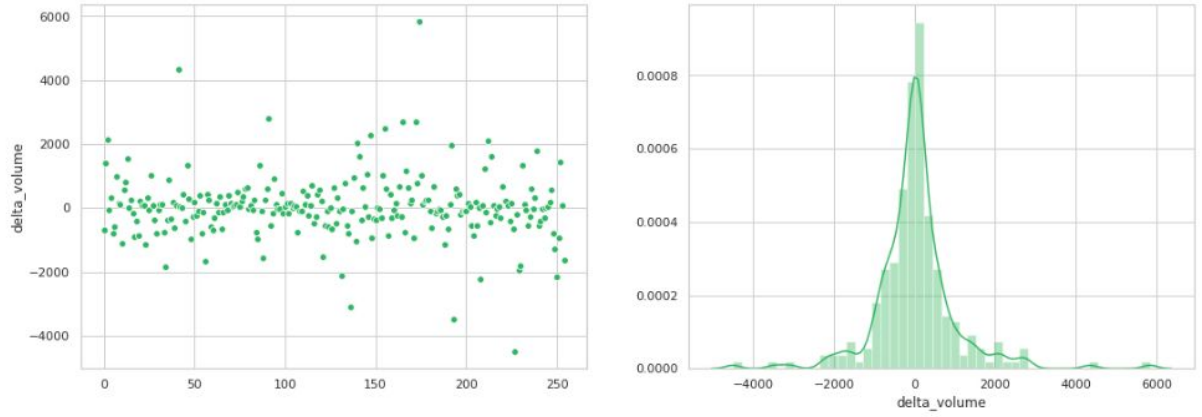
выполнил Никифоров Глеб Владиславович

Немного о точности предсказаний

В результате проделанной работы, на тестовом датасете достигнута точность в **26%** по метрике **SMAPE** и критерий **R2 = 0.87**, что, учитывая довольно незамысловатую предобработку данных и использование одной из базовых моделей регрессии, можно считать неплохим результатом.



Визуально предсказанные значения объемов продаж неплохо повторяют истинные.



Разности между истинными и предсказанными значениями имеют распределение близкое к нормальному со средним значением 33 и стандартным отклонением 1000.

О подборе оптимальной цены

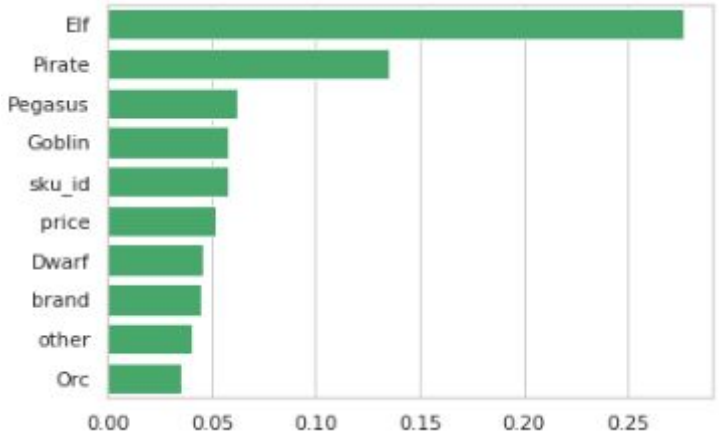
Как подбиралась цена

Оптимальная цена подбиралась для каждого товара в каждой географической области, где он продавался в июне 2019 г. Было решено, что такая гранулярность обеспечит максимальную точность. Подбор происходил перебором. Отрезок от цены-15% до цены+15% равномерно разбивался на 1000 точек, для каждого значения цены делалось предсказание обученной ранее моделью и выбиралась цена, обеспечивающая максимум произведения цены на объем продаж.

Благодаря алгоритму, позволяющему производить варьирование не для каждого товара в отдельности, а для всех сразу, перебор работает довольно быстро, не теряя при этом точности.

О достоверности

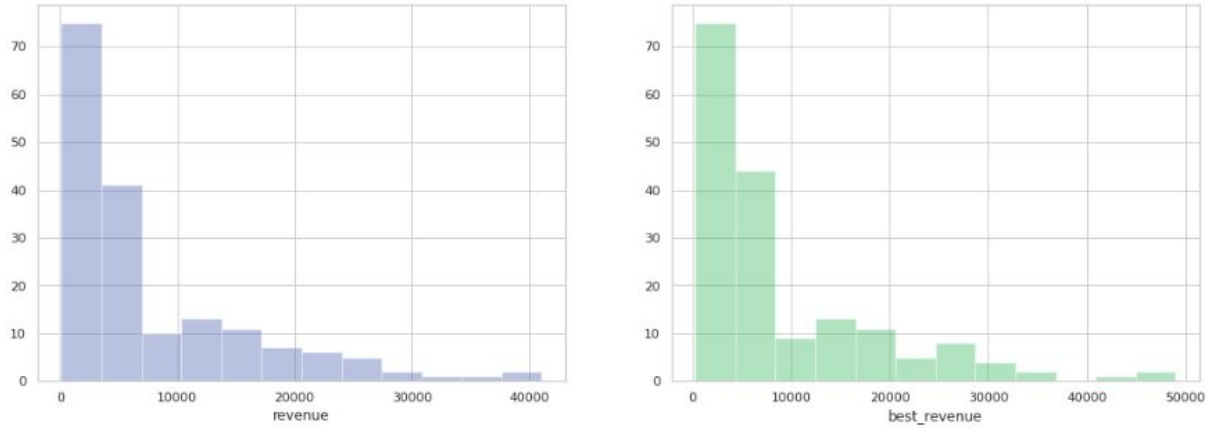
Отметим, что цена товара находится на ***шестом*** месте по важности для предсказания модели, и ее важность составляет **~5%**
Топ 10 важных признаков:



Заметим, что, судя по нашей модели, поток эльфов оказывает наибольшее влияние на объем продаж.

Было получено, что выбор оптимальной цены позволяет увеличить выручку на ***15,6%***. При этом выручка увеличивается в каждом географическом регионе для каждого товара. Была статистически проверена возможность того, что эта разница получилась случайно. Для этого использовался **t-тест**, в результате которого получено, что **p-значение** для гипотезы о случайности различий в результатах **сильно меньше 0.05**, что позволяет смело откинуть нулевую гипотезу.

Распределения выручек до и после изменения цен выглядят следующим образом:



Как можно было ожидать, они не нормальные. Однако, для больших выборок (более 30 наблюдений) нормальность распределений перестает быть необходимым условием качественного t-теста. А вот гомогенность дисперсий все-таки нужна, и она была проверена с помощью теста Левена. В результате теста было получено p-значение равное 0.33, следовательно гипотеза о гомогенности дисперсий не может быть отвергнута.

Возможности для улучшения

Одним из возможных подходов, которые могли бы увеличить точность предсказания, является использование теории временных рядов. По сути, информация для каждого товара в каждой географической зоне - это отдельный временной ряд. С помощью несложных преобразований можно сделать наши временные ряды стационарными. А теория стационарных временных рядов в свою очередь изобилует довольно точными предсказательными моделями.