Предсказание спроса в ритейле

Подготовка данных

Проведен описательный анализ публичного датасета о продажах интернет-магазина. Построены графики продаж по каждой категории товара в недельном, месячном и годовом разрезах. Для тестирования предсказательных моделей проведена группировка данных: посчитаны суммы продаж за каждый день. С целью проверки полученного ряда на стационарность проведен тест Дики-Фулера. Оказалось, что ряд стационарный, поэтому нет необходимости применять такие техники как дифференцирование или сезонное дифференцирование. В качестве тренировочных данных взяты первые 1000 наблюдений, еще 419 используются в качестве тестовых.

ARIMA

В качестве базовой модели выбрано Auto-regressive integrated moving average, одну из самых популярных статистических моделей для предсказания временных рядов, в том числе спроса в ритейле. Для выбора гиперпараметров модели построены графики автокорреляционной и частичной автокорреляционной функций: как известно, оптимальным значением p в модели AR(p) считается последний значимый пик ACF, а оптимальным значением q в MA(q) - последний значимый пик PACF.

Затем процесс подбора гиперпараметров автоматизирован с помощью библиотеки sktime. Для подбора оптимальных значений p, d, q использовался grid search - перебор всех разумных комбинаций гиперпараметров - с кросс-валидацией: модель последовательно обучалась на 70 предыдуших значениях и предсказывала 3 следующих. Предсказание на 3 дня вперед выбрано неслучайно: Для наглядности приведен пример кросс-валидации временного ряда, в котором размер тренировочной выборки на каждом шаге равен 5, а размер тестовой так же 3.

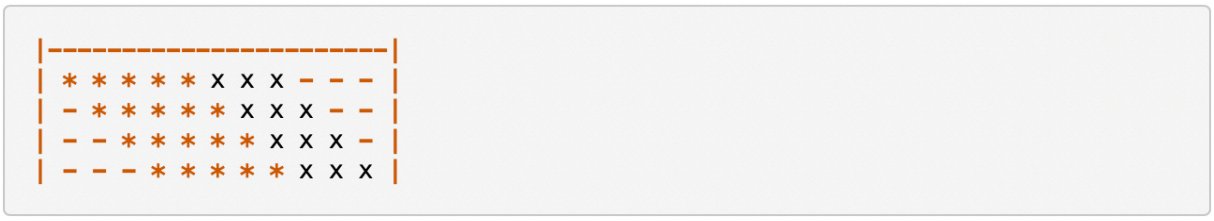

Figure 1: * - тренировочные данные, x - тестовые данные

Для сравнения результатов ARIMA с разными наборами (p, d, q) была выбрана классическая метрика Mean Absolute Percentage Error. Лучшая модель - ARIMA(3, 1, 1) с MAPE=0.31 на тестовой выборке. Таким образом, удалось добиться относительно разумных предсказаний модели.

Проведен анализ ошибок модели: достигнуты требуемые свойства “остатков” временного ряда - некореллированность, нулевое математическое ожидание. Ошибки имеют распределение, близкое к нормальному.

VAR

Вместе с тем, кажется разумным, что продажи одних товаров могут влиять на продажи других в разной степени. Как минимум, продажи той же категории товаров в предыдущие периоды влияют на продажи этой категории сегодня больше, чем продажи других категорий. Какие-то товары могут являться комплементами для других: спрос на машины приведет к увеличению спроса на бензин. С другой стороны, возможны и обратные ситуации: повышение спроса на бабл-ти окажет отрицательный эффект на продажи кофе и наоборот, потому что человек скорее всего заменяет один товар другим.Чтобы учитывать эти взаимосвязи, можно применить модель векторной авторегрессии (VAR), что и было сделано.

Данные были разделены на три основных категории, которые торгуются в данном магазине: офисные товары, технологические товары и мебель. Из исходного $y_t \in \mathbb{R}$ получился $\overline{y}_t \in \mathbb{R}^3$, к которому и применялась VAR. Подбор лучший значений был произведен с помощью grid search с использованием многопоточных вычислений для увеличения производительности.

Оказалось, что лучшая модель - это VAR(10). В большинстве случаев удалось добиться неплохих показателей по метрике MAPE, однако из-за того, что в некоторых категориях в некоторые дни продажи могут быть равны нулю, подсчет MAPE является численно неустойчивым, из-за чего в некоторых случаях получются экстремально большие значения этой метрики.