Предсказание спроса в ритейле

Подготовка данных

Проведен описательный анализ публичного датасета о продажах интернетмагазина. 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 р 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 такой подход, как кросс-валидация: модель последовательно обучалась на 70 предыдуших значениях и предсказывала 3 следующих. Для наглядности приведен пример кросс-валидации временного ряда, в котором размер тренировочной выборки на каждом шаге равен 5, а размер тестовой так же 3.

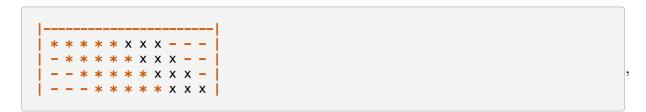


Figure 1: * - тренировочные данные, х - тестовые данные Для сравнения результатов ARIMA с разными наборами (p, d, q) была выбрана классическая метрика Mean Absolute Percentage Error. Лучшая модель - ARIMA(3, 1, 1) с MAPE=0.31 на тестовой выборке. Таким образом, удалось

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

добиться относительно разумных предсказаний модели.

\mathbf{STL}

Использован метод STL (декомпозиция по сезонным трендам с помощью локальной регрессии) с недельной периодичностью. Оказалось, что в данных есть небольшой возрастающий тренд.