Прогнозирование сезонного спроса товаров в ритейле

Песня Полина
Кафедра Анализа Данных
Научный руководитель: Владимир Кукушкин

8 марта 2022 г.

Аннотация

Основная задача данной работы – это построить модель машинного обучения для прогнозирования спроса товаров в ритейле. Ритейл – это розничные продажи конечному покупателю. Основная задача ритейла – максимизировать прибыль. Для этого необходимо уметь предсказывать пользовательский спрос на товары. Главная особенность задачи заключается в том, что товаров, представленных на маркетплейсе, и как следствие необходимых моделей для прогнозирования, очень много, порядка сотен тысяч. Нужен способ справиться с таким количеством моделей: отказаться от прогнозирования товаров с нерегулярной и плохо прогнозируемой статистикой и построить универсальную модель прогноза, которая будет достаточно хорошо работать на временных рядах с очень разными данными.

1 Постановка задачи

Цель ритейла – максимизировать прибыль – осуществляется за счёт простой схемы: купить подешевле, продать подороже. Кроме того, важным условием достижения этой цели является максимизация оборачиваемости товара: хотелось бы закупать и хранить на складе только те товары, которые бы с большей вероятностью были бы куплены. При этом есть товары, спрос на которые зависит от сезонных факторов. Например, на зимние лыжи зимой спрос будет сильно выше, чем летом, и наоборот, никому не нужны надувные круги и пляжные зонтики зимой. Поэтому необходимо разработать модель, которая бы учитывала в том числе и сезонный спрос товаров, позволяя ритейлерам оптимизировать их политику закупок и хранения на складах. Отсюда имеем:

Цель: Научиться предсказывать повышение/понижение спроса на товары, в зависимости от сезонных и других факторов.

При такой постановке мы имеем, по сути, задачу прогнозирования временного ряда.

Временной ряд – это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки. Задача прогнозирования временного ряда - это получение прогноза значений ряда на будущие моменты времени, используя его исторические данные.

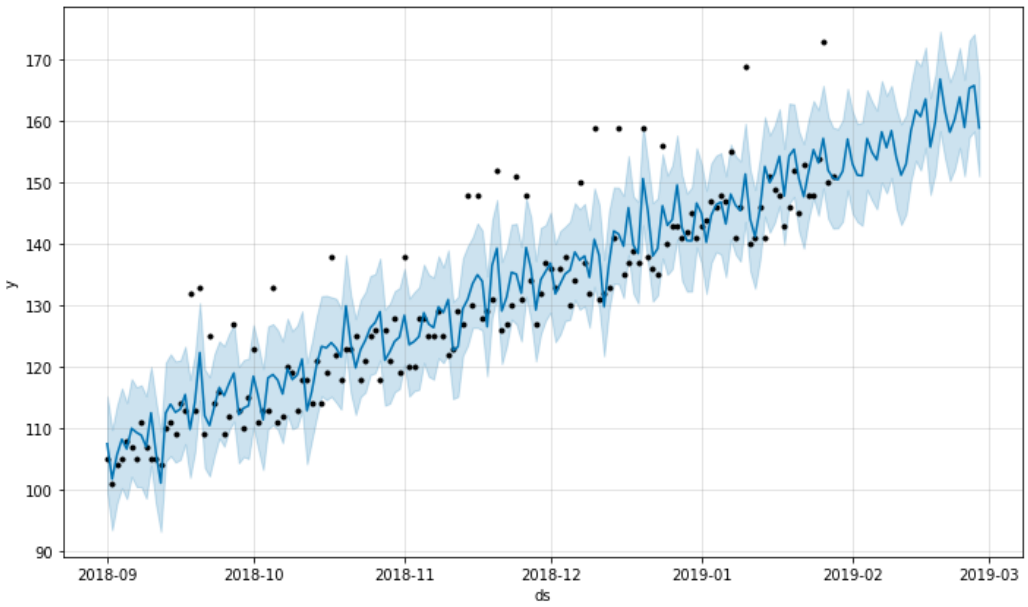

Рис. 1: Пример работы модели прогнозирования временного ряда
Чёрными точками обозначены фактические значения, на основе собранного датасета. Временной отрезок, где их нет - тестовая выборка. Синей линией обозначено предсказание модели, закрашенная голубая область обозначает 95% доверительный интервал для предсказания.

В такой постановке задача имеет множество известных решений, однако в нашем случае она имеет ряд дополнительных особенностей, усложняющих решение, а именно:

1. Нет чёткого определения сезонности.

   С одной стороны мы хотим, чтобы модель строила свой прогноз на основе данных сезонности, но с другой стороны у нас нет чёткой формулировки того, что собственно есть такое сезонность. Это важно, потому что иногда нас волнует не само предсказанное значение спроса, а его изменение относительно предыдущих значений. То есть, например, у нас может быть какой-то тренд спроса, и нужно чтобы модель в каком-то виде возвращала отклонения от этого тренда в зависимости от месяца и дня недели. Здесь существует множество различных подходов, например:

   - $forecast = trend * (seasonality_{yearly} + seasonality_{weekly} + seasonality_{holiday})$
   - $forecast = trend + seasonality + remainder$
   - ...

   В своей работе я буду придерживаться именно первого понимания сезонности

2. Товаров очень много.

Товаров, для которых мы хотим построить прогноз, очень много, порядка 10^6. Необходимых моделей, соответственно, столько же. При таком количестве невозможно руками корректировать работу каждой отдельной модели, невозможно отслеживать качество работы. Отсюда мы приходим к необходимости разработать универсальную модель, которая сможет одинаково хорошо работать с временными рядами от совершенно разных товаров. При этом, возможно, такая модель будет принимать на вход не только исторические данные, но и данные, связанные со свойствами конкретного товара.

Цель: Разработать относительно универсальную модель прогнозирования временных рядов.

3. В данных много пропусков.

   И проблема здесь не только в том что датасет собран плохо, но и в том, что в некоторых категориях товаров, как, например, паровые котлы, продажи случаются в принципе редко (см. рис. 2). Предсказывать значения такого ряда будет нецелесообразно, ввиду отсутствия нормальных исторических данных.

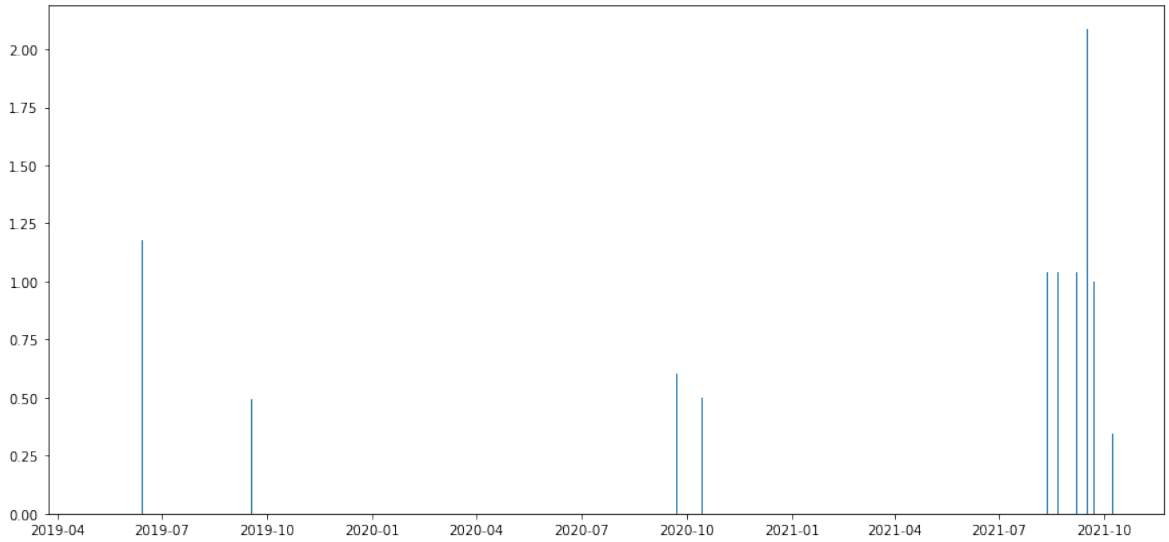

Рис. 2: Пример временного ряда с большим количеством пропусков

Впрочем, мы и не ставим целью предсказывать поведение подобных рядов. Поэтому нужно научиться искючать такого рода данные из датасета.

Цель: Определить метрику, которая бы определяла, какие временные ряды годятся для прогноза, а какие нет.

2 Данные и метрики

Данные:

Датасет состоит из исторической информации по продажам различных товаров. Строка данных содержит:

- Описание товара в порядке вложенности одного в другое (на примере смартфона): категория товара (смартфон), модель (iPhone 13 pro), msku (iPhone 13 pro 256GB space grey) – точное описание товара.
- День продажи
- GMV данной тройки (совокупность продаж в денежном выражении)

При этом различных категорий порядка 3 000, моделей - порядка 10^4, msku - порядка 10^5.

Метрики качества:

Для сравнения моделей можно использовать следующие метрики:

- Mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$$

Где e_t - ошибка в предсказании модели на конкретную дату, y_t - фактическое значение на дату t, n – длина временного ряда.

- Mean absolute scaled error

$$MASE = \frac{1}{n}\cdot\frac{\sum_{t=1}^{n} e_t}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - t_{t-m}|}$$

Это средняя абсолютная ошибка значений прогноза, деленная на среднюю абсолютную ошибку одношагового наивного прогноза в выборке. Эта безмасштабная метрика ошибок может использоваться для сравнения методов прогнозирования в одном ряду, а также для сравнения точности прогнозов между рядами.

При этом важно нормировать метрику на абсолютные значения таргета, потому что от товара к товару он сильно отличается. Продажи завязаны на стоимость товара, а потому абсолютные значения продаж для, например, продаж смартфона, будут сильно выше, чем абсолютные значения продажи продаж ластика. Без этой нормировки качество полученных моделей будет невозможно сравнить из-за разного масштаба.

3 Краткий обзор литературы

В этом разделе будет показано, что уже сделано другими исследователями в данной области.

1. ARIMA

   Это модель авторегрессии-скользящего среднего:

$$a(L)\nabla^d X_t = b(L)\varepsilon_t$$

   Где $\nabla^d X_t = (1-L)^d X_t$

$L: \ LX_t = X_{t-1}$ - оператор запаздывания.

Авторегрессионная интегрированная скользящая средняя - это одна из первых моделей для прогнозирования временных рядов. Но в нашем случае это не самое лучшее решение, потому что ARIMA требует тщательного подбора параметров для каждого отдельного временного ряда, что мы не можем себе позволить.

2. Библиотека fbprophet

   Это библиотека, разработанная Facebook для автоматического прогнозирования временного ряда. Для прогноза ряд представляется в виде:

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

   Где $g(t)$ - непериодическая составляющая (тренд). Кусочно-линейный или логистический тренд,

   $s(t)$ - периодические изменения (сезонность). Годовая - Ряд Фурье, недельная – фиктивные переменные,

   $h(t)$ - последствия праздников с нерегулярным расписанием (период – год),

   $e(t)$ - непрогнозируемая ошибка.

   Эта модель на данный момент используется в качестве **бейзлайна** задачи, над которой я работаю

3. Библиотека Darts

   Данная библиотека используется для предсказания сразу нескольких временных рядов.

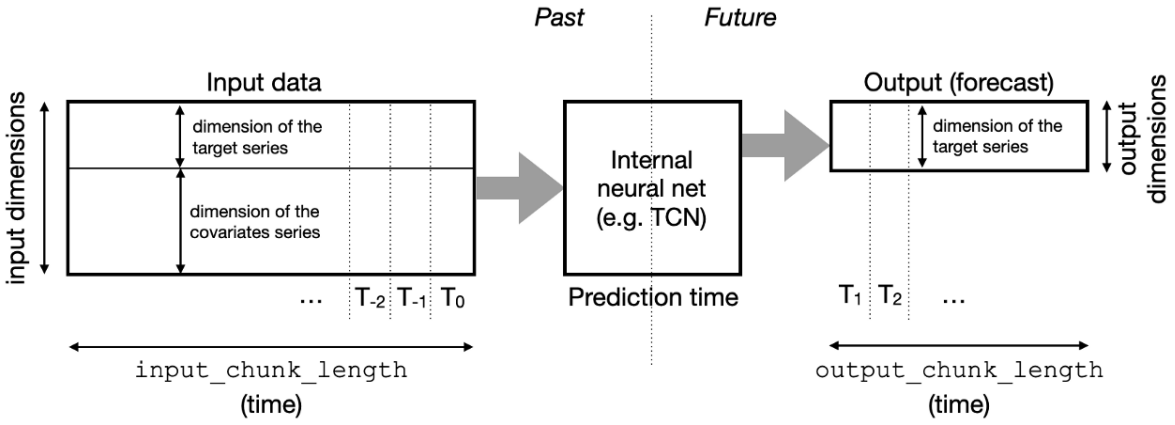

Рис. 3: Схема работы модели

Модель получает на вход несколько основных временных рядов, которые требуется предсказать, плюс несколько вспомогательных рядов, предсказывать которые не нужно, но на которые можно опираться в процессе построения модели. Далее данные пропускаются через нейронную сеть (например, рекуррентную), и на выходе получаются предсказания для основных временных рядов.

4 Уже проведённые эксперименты

1. Обучила бейзлайн модель с параметрами: мультипликативная модель сезонности, Российские праздники. Так выглядит выход работающей модели:

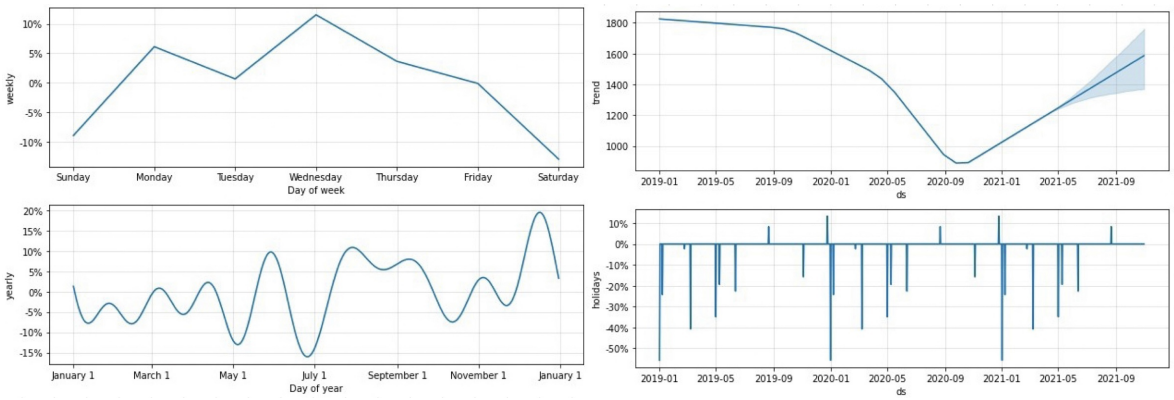

Рис. 4: Выход обученной бейзлайн модели

Модель выводит вычисленные показатели сезонности в течение года по месяцам, в течение недели по дням недели, в течение года по праздникам и прогноз на небольшой промежуток времени вперёд. Такая модель даёт такие метрики:

```
269: {'mae': 2.770624608923382, 'mape': 7.311556916772235},
270: {'mae': 4.579921681949919, 'mape': 1.132969374019734},
271: {'mae': 4.153095966792119, 'mape': 4.702190549038751},
272: {'mae': 81.64139315353863, 'mape': 145.41310745441322},
273: {'mae': 3.030631107080762, 'mape': 8.190052304068503},
274: {'mae': 594.7499302714128, 'mape': 804.3904935314362},
276: {'mae': 8.892929902840288, 'mape': 21.45443997675695},
277: {'mae': 32.678902630750194, 'mape': 303.33602516873333},
278: {'mae': 7.5721969324292795, 'mape': 13.54223138056416},
279: {'mae': 77.36369007350255, 'mape': 358.31955709485186},
280: {'mae': 7.034137829048599, 'mape': 6.368411038743515},
281: {'mae': 71.23544648952024, 'mape': 0.5737134233009552},
282: {'mae': 0.5407590109943212, 'mape': 1.3269937449446307},
283: {'mae': 24.061559551536185, 'mape': 15.168456931194157},
```

Рис. 5: Метрики бейзлайн модели

Уже по метрикам MAPE можно судить, что в некоторых случаях модель отрабатывает довольно плохо (MAPE равное единице говорит о том, что ошибка, совершаемая моделью, равна абсолютному значению таргета)

2. Провела очистку данных

   Очистка данных производилась сразу по двум критериям:

   - По отношению количества пропусков в данных (то есть дней, когда продаж не было) к длине периода всех данных. Эмпирически (чтобы в результате был отброшен адекватный процент всех данных) был подобран порог, выше которого ряды отбрасывались - то есть пропусков в данных было слишком много по отношению ко всем данным. Для каждого из трёх разбиений по

множествам: на категории товаров, на модели (мельче, чем на категории), на msku (мельче, чем на модели) порог подбирался свой. Значения порогов соответственно: 0.55, 0.97, 0.97.

- Слишком короткие ряды (одна-две точки) отбрасывались.
- Многие товары сначала плохо продаются, а потом входят в популярность, либо налаживаются закупки, и они начинают продаваться хорошо. Пример:

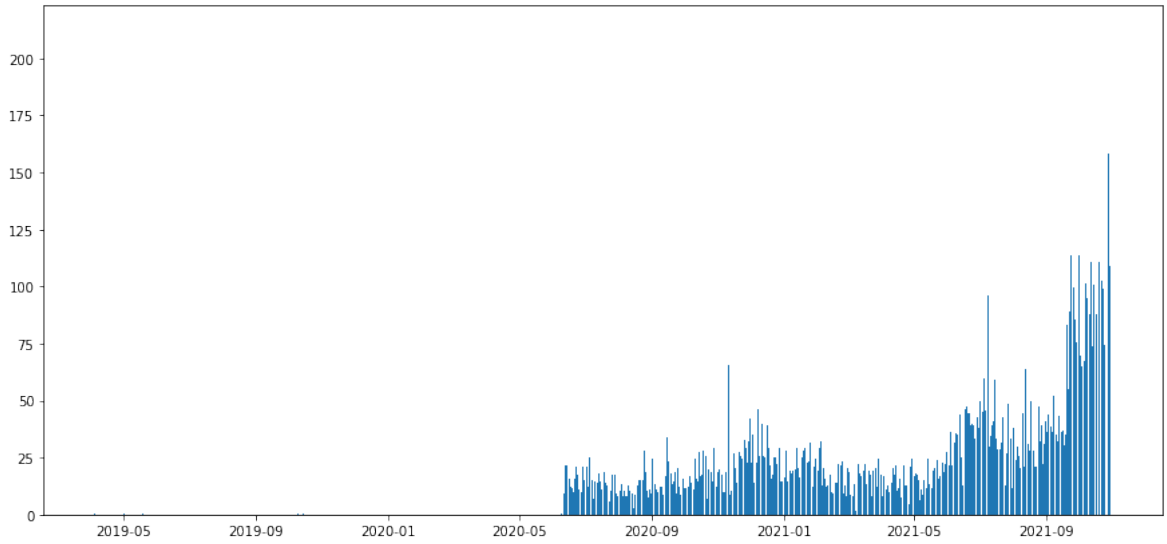

Рис. 6: Пример резко возросшего спроса

В таких рядах хочется отрезать незначительный шум в начале, тем самым упростив модели задачу, очистив историю, плюс уменьшив размер данных.

Решение: Проходимся по временному ряду с окном шириной 30 дней, считая среднее значение таргета в окне. Далее отступаем на определённое значение количества дней от конца ряда, и идя в обратном направлении ищем точку, где среднее значение в окне становится меньше $1/5$ от среднего средних в окне. По этой точке обрезаем ряд - оставляем только то, что в конце. Отступ вначале нужен, чтобы в случае, если ряд заканчивается незначительным шумом - товар уже почти перестал продаваться, не уничтожить ряд совсем.

3. Построена простая модель линейной регрессии над временным рядом, которая в качестве входных параметров получает фичи двух типов:

   - Оконные фичи

     Это фичи, которые строятся на основе окна шириной 30 дней, получая некоторые особенности последних исторических данных. Их полных список и описание можно посмотреть в разделе «Приложение» в конце.

   - Точечные фичи

     Это фичи, которые строятся только на основе одной точки. К ним относятся dummy-переменные месяца и дня недели, а так же исторические значения

курса рубля к доллару на этот день - эта фича должна помочь модели справляться с инфляцией и ростом цен на товары.

Здесь так же в качестве фичи была попытка использовать предсказания fbprophet, но это оказалось слишком затратным по времени - пол часа на сбор фичей для одного временного ряда.

Далее обученная на одном временном ряду модель получает на вход окно в 30 дней, строит по нему оконные фичи, затем делает предсказание на следующие семь дней, применяя линейную регрессию над фичами (точечные фичи строятся над точками, на которые делается предсказание). Полученные метрики MAPE на нескольких обученных рядах: 0.6 и 0.3. Что уже оказалось лучше, чем соответствующие бейзлайн модели (1.25 и 1.86 соответственно)

5 Дальнейшие планы

В дальнейшем планируется:

1. Придумать способ как распространить несколько моделей, обученных на временных рядах на остальные временные ряды. В качестве возможного решения можно добавить в модель фичи, которые связывают её с другими товарами, которые находятся в той же категории товара, а потом распространить решение на всю категорию.
2. Придумать метрики сравнения и сравниться с бейзлайн моделью

6 Литература

1. George E. P. Box, Gwilym M. Jenkins.Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics). 1970
2. Sean J. Taylor, Benjamin Letham (Facebook). Forecasting at Scale. [Online], 2017
3. Julien Herzen. Training Forecasting Models on Multiple Time Series with Darts. [Online], 2021

7 Приложение

Полный список оконных фичей – фичей, построенных на окне шириной в 30 дней – используемых в работе:

1. Среднее в окне – среднее арифметическое всех значений в окне.
2. Среднее во второй половине окна – среднее арифметическое от последних 15 значений окна.
3. Среднее в последней четверти окна – среднее арифметическое от последних 7 значений окна.

4. Максимум в окне – максимальное значение среди всех значений окна.
5. Минимум в окне – минимальное значение, среди всех значений окна.
6. Дисперсия в окне – дисперсия всех значений окна.
7. Последние три значения окна – каждое из трёх значений используется в качестве отдельной фичи.

Все приведённые выше фичи также строятся на ряде приростов – ряде, полученном из исходного взятием разности между двумя соседними значениями ряда. То есть такой ряд показывает буквально прирост значения продаж за сутки. Всего получается 14 оконных фичей (7 на исходном ряде, и 7 на ряде приростов)