Введение

Спецификой ритейла является высокая волатильность спроса и сильно ограниченное время реализации товаров [1]. В этой связи прогнозирование спроса является критически важной задачей для бизнеса, для решения которой используются методы статистики, машинного обучения и нейронных сетей [2].

Прогнозирование временных рядов, в частности задача предсказания спроса, является классической задачей регрессии. Требуется найти отображение, которое данному временному ряду - последовательности $y_1, ..., y_t$ сопоставит набор предсказаний - $\overline{y}_{y+1}, ..., \overline{y}_{t+H}$, минимизировав разницу между предсказаниями и настоящими значениями $y_1, ..., y_{t+H}$

Задачу предсказания временных рядов решают разными способами. Во-первых, это "модели временных рядов", предполагающие различные виды линейной зависимости между членами временного ряда. Во-вторых, это методы машинного обучения, такие как линейная регрессия, решающие деревья, kernel regression, градиентный бустинг. Наконец, в последние годы для предсказания временных рядов стали активно использоваться нейросети (aritificial neural networks) [1].

Особенностью спроса как временного ряда является то, что фирма может влиять на некоторые из признаков, задающих спрос, а именно на признаки связанные с ценой. В этой связи методы прикладной математики активно применяются для решения задач оптимизации, свзяанных с максимизацией прибыли при известной функции спроса. В честности, модели спроса основанные на эластичности по цене (elasticity based demand function, EDF) [7].

Поскольку задача непосредственно связана с выгодой для бизнеса, открытых исследований на эту тему немного: компании, продающие свои консультационные услуги по выставлению оптимальных цен, не заинтересованы в открытости методов вычисления эластичностей спроса.

ARIMA

В качестве базовой модели выбрано Auto-regressive integrated moving average, одну из самых популярных статистических моделей для предсказания временных рядов [6]. ARIMA(p, d, q) предполагает следующую зависимость между членами временного ряда:

$$\begin{gathered} y_t = \alpha_1 y_{y-1} + \ldots + \alpha_p y_{t-p} \\ +\beta_1 \varepsilon_1 + \ldots + \beta_q \varepsilon_{t-q} + \varepsilon_t \end{gathered} \tag{1}$$

где y_t - значение в момент t, ε_i - “белый шум” (компонента ряда, которую невозможно предсказать) в момент t

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. Данная модель также широко применяется для решения поставленной задачи [2, 3].

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 [4].

Постановка задачи

Неформальная постановка задачи

Как уже было сказано, задача регрессии успешно решается различными способами (часть из которых описана выше). Вместе с тем, если рассматривать процесс деятельности ритейлера не со стороны наблюдателя, а со стороны ритейлера, возникает более общая задача - задача максимизации прибыли. Обучая модель регрессии, мы получаем ответ на вопрос: как зависит спрос от различных факторов, которые на него влияют? Среди этих факторов - цена на товар, которую на самом деле фирма устанавливает сама. Поэтому с точки зрения ритейлера можно говорить не только о поиске отображения из пространства признаков в пространство объектов, но и о поиске оптимального значения цены, которое фирме стоит выставить.

Формальная постановка задачи

Для начала требуется решить задачу регрессии: по данной выборке $X \in \mathbb{R}^{n \times m}$ построить отображение $f : \mathbb{R}^n \to \mathbb{R}$, которое приближает функцию спроса $y(x), x \in \mathbb{R}^n$, минимизируя функцию потерь. В нашем случае в качестве loss function $\mathcal{L}(X; \omega, b)$ выбрана MAPE - mean absolute mercentage error. Зависимость спроса от признаков предполагается линейной:

$$y = \omega X + b \tag{2}$$

И задача заключается в

$$\mathcal{L}(X; \omega, b) \to \min \omega \in \mathbb{R}^n, b \in \mathbb{R} \tag{3}$$

где

$$\mathcal{L}(X; \omega, b) = \frac{1}{m} \Sigma \, |\frac{\hat{x_i} - x_i}{x_i}| \tag{4}$$

Далее для простоты иногда будем пропускать b, потому что свободный член можно без ограничения общности воспринимать как коэффициент для признака, значения которого равно 1 у всех объектов выборки.

Введем следующие обозначения:

$$
\begin{gathered}
p_t - \text{цена товара в период } t, \\
\Delta_p = \frac{p_t - p_{t-1}}{p_{t-1}}, \\
T - \text{длина сезона для сезонной модели} \\
a \in \mathbb{R}^n - \text{все остальные признаки, на которые фирма не может влиять,} \\
\mathrm{TR}_t = p_t * y_t, \\
\mathrm{TC}_t - \text{издержки фирмы,} \\
\pi = \mathrm{TR} - \mathrm{TC}
\end{gathered} \tag{5}
$$

Требуется решить задачу оптимизации:

$$
\pi \to \max \Delta_p \tag{6}
$$

Предлагаемая модель

Описание модели

На текущий момент большинство исследователей, моделирующих спрос с помощью линейных моделей, предполагают используют аппарат линейной регрессии. В этой работе предлагается использовать аппарат интегрированной модели авторегрессии-скользящего среднего (ARIMA) для моделирования спроса.

Базовая персия модели рассматривает спрос Y_t как временной ряд, значение которого зависит от признаков, которые можно разделить на несколько групп:

- лаговые признаки (auto-regressive features)
- ошибки модели в предыдущие периоды (moving-average features)
- сезонность
- остаточный компонент (residuals)
- Δ_p - изменение цены

В виде формулы это можно изобразить в следующем виде:

$$
\begin{gathered}
y_t = \alpha_1 y_{t-1} + \dots + \alpha_k y_{t-k} \\
+ \beta_1 \varepsilon_{t-1} + \dots + \beta_{t-m} \varepsilon_{t-m} \\
+ \gamma_1 y_{y-T} + \dots + \gamma_l y_{t-lT} \\
+ \alpha \Delta_p + \beta \varepsilon_t
\end{gathered} \tag{7}
$$

С использованием обозначений выше:

$$y_t = \alpha\Delta_p + \beta\varepsilon_t + a^T x \tag{8}$$

Таким образом, модель можно представить в виде $\mathrm{SARIMA}(p, d, q, (P, D, Q)_T)$ с экзогенной переменной Δ_p. Такую модель также называют SARIMAX - сезонная ARIMA с вектором x в качестве экзогенной переменной [8].

Задача оптимизации

Так как оптимизация производится по Δ_p, издержки фирмы являются константой, поэтому достаточно максимизировать величину

$$\mathrm{TR} = p_{t-1}(1 + \Delta_p) \times y_t(\Delta_p) \tag{9}$$

Так как функция спроса предполагается линейной (см. Equation 8), решение задачи оптимизации выписывается явным образом:

$$\begin{gathered}\frac{\partial\ \mathrm{TR}}{\partial\Delta_p} = p_{t-1}y_t(\Delta_p) + p_{t-1}(1 + \Delta_p)y_t'(\Delta_p) = \\ = p_{t-1}y_t(\Delta_p) + \alpha p_{t-1}(1 + \Delta_p) = 0 \\ \Longrightarrow a^T x + \alpha\Delta_p + \alpha(1 + \Delta_p) = 0 \\ \Longrightarrow \Delta_p = \frac{-a^T x - \alpha}{2\alpha}\end{gathered} \tag{10}$$

В случае нелинейной функции спроса, можно применять численные методы оптимизации, например, градиентный спуск.

Применение модели

Подбор гиперпараметров

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 [5].

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.

Для наглядности приведен пример кросс-валидации временного ряда, в котором размер тренировочной выборки на каждом шаге равен 5, а размер тестовой так же 3.

```
|---------------------|
| * * * * * x x x - - - |
| - * * * * * x x x - - |
| - - * * * * * x x x - |
| - - - * * * * * x x x |
```

,

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.

Источники

Список литературы

[1] Da Veiga CP, Da Veiga CR, Catapan A, Tortato U, Da Silva WV. Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. WSEAS transactions on business and economics. 2014 Jan;11(1):608-14.

[2] Tsolacos S. Econometric modelling and forecasting of new retail development. Journal of Property Research. 1998 Jan 1;15(4):265-83.

[3] Brooks C, Tsolacos S. Forecasting models of retail rents. Environment and Planning A. 2000 Oct;32(10):1825-39.

[4] Hyndman RJ. Forecasting: principles and practice. OTexts; 2018. URL: https://otexts.com/fpp2/accuracy.html

[5] Robert Nau, Identifying the numbers of AR or MA terms in an ARIMA model, Duke University, 2020 URL: https://people.duke.edu/~rnau/411arim3.htm

[6] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed). Hoboken, New Jersey: John Wiley & Sons.

[7]Elasticity Based Demand Forecasting and Price Optimization for Online Retail, Chengcheng Liu, M´aty´as A. Sustik Walmart Labs, San Bruno, CA, June 17, 2021. URL: https://arxiv.org/pdf/2106.08274

[8] Vagropoulos SI, Chouliaras GI, Kardakos EG, Simoglou CK, Bakirtzis AG. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In2016 IEEE international energy conference (ENERGYCON) 2016 Apr 4 (pp. 1-6). IEEE.

Приложения

[1] Репозиторий проекта
URL: https://github.com/chagrygoris/retail_forecasts