МОДЕЛИ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ НАГРУЗКИ РАБОТНИКОВ СЛУЖБЫ УПРАВЛЕНИЯ ПЕРСОНАЛОМ

А. С. Подгорный

Дальневосточный федеральный университет, Владивосток, Россия
E-mail: andreypodgorny10@gmail.com

В настоящей статье рассмотрено применение различных технологий прогнозирования временных рядов на примере данных о задачах, поступающих в службу управления персоналом через автоматизированные информационные системы. Модель прогноза ожидаемой нагрузки на работников позволяет скорректировать ресурсы на горизонт прогнозирования. Были рассмотрены различные технологии анализа временных рядов с целью выбора модели, показавшей наилучшее качество на имеющихся данных.

TIME SERIES FORECASTING MODELS FOR HR RESOURCES LOAD PREDICTION

A. S. Podgornyi

This article describes the use of various time series forecasting technologies for the data of tasks received by Human Resources department via automated information systems. The forecast model of the expected employees load help to adjust resources for the forecast horizon. Various time series analysis technologies were considered in order to select the model that showed the best quality on the available data.

В крупной российской компании, сотрудники которой распределены по всей территории страны и в разных часовых поясах, реализован централизованный сервис обработки задач, поступающих от работников компании в адрес службы управления персоналом. Централизация сервиса помогает экономить бюджет, однако, в случае резкого увеличения потока входящих задач в сравнении с доступным ресурсом, может привести к реализации риска перегрузки работников службы и к невыполнению задач в срок. Специфика заключается в нерегулируемости входящего потока задач под ограничение доступного ресурса.

Исходя из постановки задачи заказчиком, нас интересует только прогноз случаев превышения допустимого максимума нагрузки, но для комплексной оценки качества модели указывать такую метрику как оптимизационную при построении было бы некорректно, ввиду упущения случаев прогноза, когда нагрузка является допустимой. В связи с этим использованы классические метрики для моделей временных рядов: MAE, RMSE, WAPE. Нами было рассмотрено применение классических моделей временных рядов, такие как ARIMA, GARCH, Silverkite, модели бустинга и экспоненциального сглаживания.

Основной задачей является разработка модели, которая сможет легко настраиваться и при этом быстро работать. При этом важно наличие настроек для интерпретируемости и возможность вода особых данных пользователем и учета их моделью (например, особых событий, аномалий, праздников). В части интерпретируемости важно помочь пользователям исследовать и валидировать модели прогнозирования с помощью экспертных знаний. В том числе, прогноз модели нужно разложить на различные компоненты (например, на сезонность, события, тренд) для их валидации на стороне пользователя – с одной стороны, и анализа динамики – с другой.

Лучшей из показавших себя моделей стала технология Silverkite [4], которая позволяет разбить задачу моделирования на две фазы: модель условного среднего и модель волатильности (ошибки) [3]. В рамках первой фазы мы обучаем модель и используем ее для прогнозирования интересующей нас метрики (нами использовались такие метрики как MAE, RMSE, средняя процентная ошибка MAPE), а в рамках второй фазы обучаем модель волатильности на остатках.

В рамках бизнес контекста особенно важным свойством примененной модели должна была стать ее гибкость, позволяющая учитывать сложные паттерны при рассмотрении интересующего события. Гибкость необходима для достижения высокой точности. Например, на располагаемых данных динамика поступления задач в выходные дни часто отличалась от будней.

Наличие отдельной модели волатильности позволяет добиться высокой скорости выполнения в производственных задачах, где может потребоваться обновление прогноза для большого количества временных рядов [1]. Кроме того, это помогает избежать таких проблем, как расхождение моделируемых рядов, которые являются общей проблемой при использовании интегрированных моделей [2]. Скорость обучения модели может также является ключевым фактором, поскольку мы можем попробовать большое количество моделей в ходе поиска оптимальных компонент и параметров (например, речь может идти о поиске оптимального порядка рядов Фурье для рядов разной частоты и оптимальной сложности авторегрессионной компоненты).

Для моделирования сезонности (компонента условного среднего) в первой компоненте модели о прогнозировании условного среднего использовались ряды Фурье [4]. При этом следует отметить, что сезонность может проявляться на различных уровнях временного ряда. Например, для почасовых данных, как в нашем случае, можно ожидать паттерн сезонности в течении дня, паттерн недельной сезонности, месячной и квартальной. На каждом из таких уровней сезонности применяется отдельное моделирование с помощью ряда Фурье (меняется соответствующий порядок ряда):

$$S_k = \sin\big(k\omega_d d(t)\big), c_k = \cos\big(k\omega_d d(t)\big), \omega_d = \frac{2\pi}{24},$$
$$k = 1, \dots , K$$

где $d(t)$ обозначает время дня, рассмотренное выше; а K – соответствующий порядок ряда Фурье, который можно задать на этапе выбора модели. Обратите внимание, что частота ω_d установлена равной $\frac{2\pi}{24}$, поскольку время дня изменяется от 0 до 24. Аналогичным образом мы можем определить соответствующий ряд Фурье для недельной, месячной, квартальной и годовой сезонности.

Для моделирования роста (компонента условного среднего) требуется ввести некоторые базовые функции и работаем кусочно-непрерывными вариантами этих функций. Пусть t обозначает непрерывное время, представленное исходными признаками. Затем рассмотрим следующие базисные функции [3]:

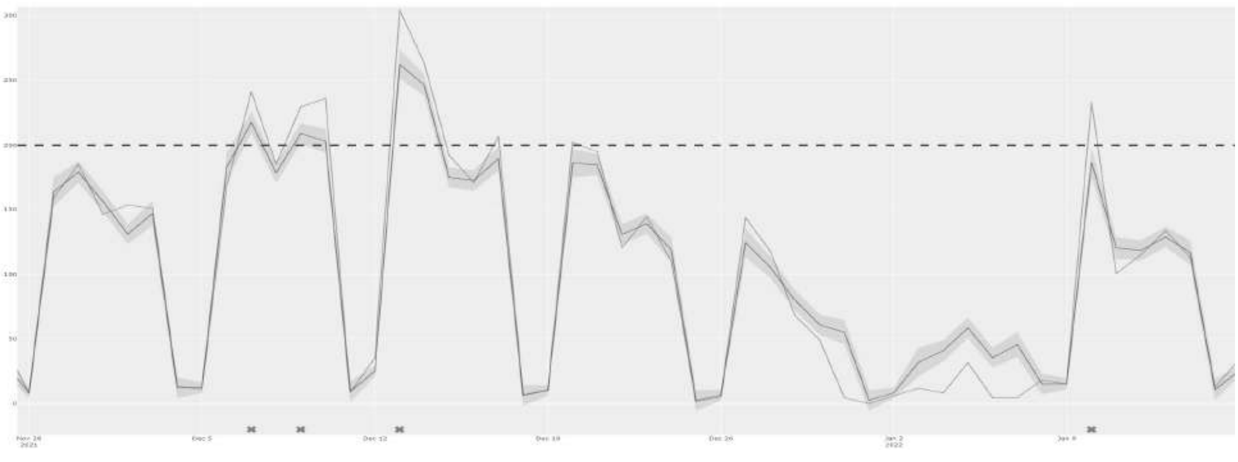
$$f(t) = t^p, p = \frac{1}{3}, \frac{1}{2}, 1,2,3.$$

Предлагается непрерывное изменение роста во времени в заданных точках изменения $t_1, \ldots, t_k$. Учитывая точки изменения и базисную функцию $f(t)$, определим

$$growth(t) = a_0 f(t) + \sum_{i=1}^{k} a_i 1_{t>t_i}(f(t) - f(t_i))$$

Следует обратить внимание, что $growth(t)$ является непрерывной функцией от времени t, но при этом производная функции может изменяться в точках изменения [5].

На рисунке ниже изображена часть графика прогноза нагрузки входящих задач: коридор прогнозов (синяя линия) сравнивается с фактическими данными (красная линия) для периода. Ось абсцисс – временная шкала, ординат – количество часов входящей нагрузки. Для удобства восприятия представлена дневная агрегация нагрузки. Зеленой пунктирной линией обозначена максимальная емкость (ресурс) подразделения по управлению персоналом. Прогнозы выше линии максимального ресурса считаются превышениями возможной нагрузки и обозначены красными крестами (ниже нуля по оси ординат).



Пример применения двухфазной модели временных рядов
для задачи прогнозирования нагрузки работников

Для качественного и масштабируемого прогнозирования в производственной среде требуется использовать гибкий фреймворк. Решение эмпирической задачи прогноза нагрузки на работников службы управления персоналом показало лучшую применимость технологии Silverkite в сравнении с классическими моделями. При этом предложенный при решении задачи дизайн продукта должен помочь создать гибкие и интерпретируемые прогнозы, а также оценки волатильности.

СПИСОК ЛИТЕРАТУРЫ

1. *Hosseini, Reza and Yang, Kaixu and Chen, Albert and Patra, Sayan* A flexible forecasting model for production systems, arXiv. 2021.
2. *Hosseini, R., Takemura, A. and Hosseini A.* Non-linear time-varying stochastic models for agroclimate risk assessment // Environmental and Ecological Statistics. 2015. Vol. 22 (2). P. 227-246.
3. *Brockwell, P., J., and Davis, R., A.* Introduction to Time Series and Forecasting // Springe. 2016. Vol. 3.
4. *Fokianos K.* Count Time Series Models Handbook of Statistics, Time Series Analysis: Methods and Applications Edited by Rao, T. S., Rao, S. S. and Rao C.R. 2012. Vol. 30. Chapter 12. pp. 315-348.
5. *West M and Harrison J.* Bayesian Forecasting and Dynamic Models, 2nd Edition. 1997.