

УДК 519.246.8

Мирзянов Тимур Русланович, магистрант, 2 курс,
Институт компьютерных технологий и защиты информации,
ФГБОУ ВО «КНИТУ-КАИ»

г. Казань

Валитова Наталья Львовна, доцент кафедры ПМИ,
кандидат технических наук,
ФГБОУ ВО «КНИТУ-КАИ»

СРАВНИТЕЛЬНЫЙ АНАЛИЗ КЛАССИЧЕСКИХ И МАШИННО- ОБУЧАЕМЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ НАГРУЗКИ НА ВЕБ- СЕРВИС

Аннотация

Рассматривается задача среднесрочного прогнозирования нагрузки на веб-сервис интернет-магазина на горизонте одной недели. Нагрузка измеряется числом пользовательских событий в час. Такие прогнозы используются при планировании вычислительных и операционных ресурсов: виртуальных машин, балансировщиков, пропускной способности баз данных. Цель работы — сравнить четыре метода прогнозирования временного ряда на едином открытом датасете и дать рекомендации по выбору модели. Сопоставляются два классических статистических метода и два метода машинного обучения, обучаемых на лаговых и календарных признаках. Эксперимент выполнен на открытом датасете Retailrocket, содержащем около 2,7 млн событий

пользовательской активности реального интернет-магазина (май–сентябрь 2015 года).

Abstract

The paper considers medium-term forecasting of an e-commerce web service workload at a one-week horizon. The workload is measured as the hourly number of user events. Such forecasts are used in planning computational and operational resources: virtual machines, load balancers, and database throughput. The aim of the work is to compare four time series forecasting methods on a single open dataset and to provide recommendations for model selection. Two classical statistical methods and two machine learning methods trained on lag and calendar features are compared. The experiment is performed on the open Retailrocket dataset containing about 2.7 million user activity events of a real online store (May–September 2015).

Ключевые слова: прогнозирование временных рядов, нагрузка на информационную систему, веб-сервис, интернет-магазин, SARIMA, метод Хольта–Уинтерса, градиентный бустинг, случайный лес.

Keywords: time series forecasting, information system workload, web service, e-commerce, SARIMA, Holt–Winters method, gradient boosting, random forest.

Введение

Интернет-магазин представляет собой информационную систему, обслуживающую поток пользовательских обращений: запросы каталога, просмотры карточек товаров, добавления в корзину, оформление заказов. Нагрузка — совокупная интенсивность этих обращений, которая характеризуется нестационарностью: помимо общего тренда, отражающего рост или убыль аудитории, в ней присутствуют две регулярные сезонные компоненты: суточная и недельная. Также присутствуют менее регулярные всплески, связанные с маркетинговыми событиями.

Среднесрочный прогноз нагрузки на горизонте одной–двух недель служит ключевым входом для задачи планирования ресурсов. По прогнозу

нагрузки заблаговременно резервируется инфраструктура: виртуальные машины, балансировщики, пропускная способность баз данных и операционные мощности. Качество прогноза определяет два экономически противоположных риска: к нехватке ресурсов и потере клиентов, к избыточным затратам на инфраструктуру.

Цель работы — сравнить классические и машинно-обучаемые методы прогнозирования временных рядов на задаче среднесрочного прогноза нагрузки веб-сервиса интернет-магазина.

Задачи:

1. Сформулировать задачу прогнозирования нагрузки в терминах теории временных рядов и обосновать выбор моделирующих средств;
2. Описать шесть рассматриваемых моделей в единых обозначениях;
3. Провести вычислительный эксперимент на открытом датасете реального интернет-магазина;
4. Сопоставить модели по метрикам MAE, RMSE, MAPE и сформулировать выводы.

1. Постановка задачи

Пусть веб-сервис принимает поток пользовательских событий. Шаг агрегации — один час. Определим временной ряд

$$y_t \in \mathbb{Z}_{\geq 0}, \quad t = 1, 2, \dots, N, \quad (1)$$

где y_t — число пользовательских событий в час t .

Задача среднесрочного прогнозирования: по наблюдаемой истории $y_1, \dots, y_N$ построить точечный прогноз на горизонте $H = 168$ часов (одна неделя)

$$\hat{y}_{N+1}, \hat{y}_{N+2}, \dots, \hat{y}_{N+H}, \quad (2)$$

Качество оценивается тремя стандартными метриками. Обозначим ошибку прогноза в момент t как $e_t = y_t - \hat{y}_t$. Используются: средняя абсолютная ошибка MAE, среднеквадратичная ошибка RMSE и средняя абсолютная процентная ошибка MAPE. Чем меньше метрика, тем точнее прогноз. Определения метрик приведены в [1, 2].

2. Математические модели

2.1 Метод Хольта-Уинтерса

Метод Хольта-Уинтерса (тройное экспоненциальное сглаживание) представляет ряд как сумму трех медленно меняющихся компонент: уровня ℓ_t , тренда b_t и сезонной компоненты S_t с периодом s . На каждом шаге компоненты обновляются взвешенным усреднением прежнего значения с новой оценкой, полученной по наблюдению y_t [1, 2]:

$$\ell_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)(\ell_{t-1} + b_{t-1}), \quad (3)$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)b_{t-1}, \quad (4)$$

$$S_t = \gamma(y_t - \ell_t) + (1 - \gamma)S_{t-s}. \quad (5)$$

Параметры сглаживания $\alpha, \beta, \gamma \in [0,1]$ управляют «скоростью забывания». Их подбирают автоматически — минимизацией суммы квадратов ошибок прогноза на обучающей выборке. Прогноз на h шагов вперед:

$$\hat{y}_{N+h} = \ell_N + h \cdot b_N + S_{N+h-s[h/s]}. \quad (6)$$

В работе использован $s = 24$ (суточная сезонность), аддитивная форма всех компонент. Реализация с использованием библиотеки statsmodels [7].

2.2 Модель SARIMA

SARIMA (Seasonal AutoRegressive Integrated Moving Average) — обобщение классической ARIMA на ряды с сезонностью [2]. Базовая модель ARMA связывает текущее значение ряда с предыдущими значениями (авторегрессия, AR) и предыдущими ошибками прогноза (скользящее среднее, MA):

$$y_t = c + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t, \quad (7)$$

где c — константа, ϕ_i — коэффициенты AR, θ_j — коэффициенты MA, ε_t — белый шум.

Модель SARIMA расширяет классическую модель двумя механизмами:

1. Дифференцированием — для устранения тренда;
2. Сезонными слагаемыми — аналогичными AR- и MA-членами с лагами, кратными длине сезона s .

Полная модель обозначается $SARIMA(p, d, q)(P, D, Q)_s$, где p, q — порядки несезонных AR и MA, d — порядок обычного дифференцирования, P, Q, D — порядки сезонных AR, MA и сезонного дифференцирования, s — длина сезона.

Коэффициенты модели оцениваются методом максимального правдоподобия [2]. Порядки (p, d, q, P, D, Q) подбираются автоматически по информационному критерию Акаике:

$$AIC = -2\ln L + 2k, \quad (8)$$

где L — максимум правдоподобия, k — число параметров. Чем меньше AIC, тем лучше компромисс между точностью и сложностью модели. Использованы библиотеки statsmodels и pmdarima [7].

2.3 Схема методов машинного обучения

Методы машинного обучения переформулируют задачу прогнозирования как задачу регрессии. Для каждого момента t строится вектор признаков $x_t \in \mathbb{R}^p$, описывающий «контекст» этого момента, и модель учится предсказывать y_t по x_t :

$$\hat{y}_t = f(x_t; \theta), \quad (9)$$

где f — выбранный тип моделей, θ — параметры, подбираемые на обучающей выборке.

2.4 Случайный лес

Метод построен на ансамбле из K деревьев решений. Каждое дерево обучается на случайной подвыборке обучающего множества и случайном подмножестве признаков. Итоговый прогноз — среднее предсказаний всех деревьев [8]:

$$\hat{y}_t = \frac{1}{K} \sum_{k=1}^K T_k(x_t), \quad (10)$$

где $T_k(x_t)$ — предсказание k -го дерева. Каждое дерево разбивает пространство признаков на области и в каждой области предсказывает выборочное среднее y по обучающим объектам этой области. Усреднение по K независимо обученным деревьям снижает дисперсию.

Гиперпараметры: число деревьев K , максимальная глубина каждого дерева, минимальное число объектов в листе. Реализация с использованием scikit-learn [9].

2.5 Градиентный бустинг

В отличие от случайного леса, в градиентном бустинге деревья строятся последовательно. Каждое новое дерево обучается аппроксимировать остатки текущего ансамбля [10]. После M итераций:

$$\hat{y}_t = \sum_{m=1}^M \eta \cdot T_m(x_t), \quad (11)$$

где $\eta \in (0,1]$ — скорость обучения, T_m — m -е дерево. На шаге t дерево T_m обучается на парах $(x_t, r_t^{(m)})$, где $r_t^{(m)}$ — остаток предыдущего ансамбля. Малое η (обычно 0,01–0,1) делает обучение плавнее и снижает риск переобучения.

Гиперпараметры: число деревьев M , максимальная глубина, скорость обучения η , число листьев, коэффициенты регуляризации. Реализация с использованием библиотеки LightGBM [11].

3. Эксперимент

3.1 Исходные данные

В качестве исходных данных взят открытый датасет Retailrocket E-commerce Dataset [12]. Он содержит историю взаимодействий пользователей с реальным интернет-магазином за период с мая по сентябрь 2015 года. Файл events.csv содержит около 2,7 млн записей вида (timestamp, visitorid, event, itemid, transactionid). Поле event принимает три значения: просмотр товара, добавление в корзину, оформление покупки. Каждая запись соответствует

одному пользовательскому обращению к серверу и является элементом нагрузки на ИС.

Из записей сформирован часовой временной ряд y_t событий за час. Для эксперимента взят последний полный месяц (720 наблюдений).

Разбиение train/test хронологическое: последние 168 часов (одна календарная неделя) формируют тестовую выборку, остальные 552 часа формируют обучающую.

Сезонная декомпозиция ряда подтверждает наличие выраженной суточной сезонности: пик днем и вечером, минимум в раннее утро. Также менее выраженной недельной сезонности.

3.2 Параметры моделей

Во всех моделях использованы типовые значения гиперпараметров, рекомендованные в руководствах по соответствующим библиотекам:

1. SARIMA(2,1,1)(1,1,1)₂₄— порядки выбраны как разумный шаблон для часовых рядов с суточной сезонностью; коэффициенты ϕ_i , θ_j оцениваются методом максимального правдоподобия;
2. Холт–Уинтерс с аддитивной сезонностью, $s = 24$; параметры сглаживания α, β, γ подбираются автоматически минимизацией суммы квадратов ошибок;
3. Random Forest: 300 деревьев, максимальная глубина 10;
4. LightGBM: 500 деревьев, скорость обучения $\eta = 0,05$, максимальная глубина 6.

Все модели реализованы на Python с использованием библиотек statsmodels, scikit-learn и lightgbm. Прогноз ML-моделей строился прямым способом — лаги формировались из истории ряда без рекурсии.

3.3 Результаты

Сравнение моделей по метрикам MAE, RMSE и MAPE на тестовой неделе приведено в таблице.

Таблица 1. Результаты эксперимента

Метод	MAE	RMSE	MAPE, %
Хольт-Уинтерс	234,08	306,62	48,27
SARIMA	212,68	282,21	45,34
Случайный лес	137,37	199,22	26,59
LightGBM	127,28	186,55	24,98

3.4 Вывод

Метод Хольта–Уинтерса явно моделирует тренд и одну сезонность (суточную), но не учитывает автокорреляционную структуру ряда. В результате наибольшая ошибка среди методов.

SARIMA превосходит метод Хольта–Уинтерса на 9 % по MAE за счет учета авторегрессии и сезонного дифференцирования. Это лучшая из классических моделей в эксперименте.

Случайный лес обходит SARIMA по всем метрикам, MAE снижена на 35 %. Объяснение в нелинейных взаимодействиях между лагами и календарными признаками.

Градиентный бустинг показал лучший результат. Последовательное обучение деревьев на остатках дает более точный прогноз, чем независимое усреднение.

Заключение

Проведено сравнительное исследование четырех методов прогнозирования временных рядов на задаче среднесрочного прогноза нагрузки веб-сервиса интернет-магазина с горизонтом 168 часов. В сравнение включены два классических метода (Хольт–Уинтерс, SARIMA) и два метода машинного обучения (случайный лес, градиентный бустинг LightGBM). Эксперимент выполнен на открытом датасете Retailrocket — около 2,7 млн событий пользовательской активности.

Методы машинного обучения показали существенно более точный прогноз, чем классические статистические модели. Наилучший результат — у градиентного бустинга (LightGBM): он опередил SARIMA и Хольта–Уинтерса по всем трем метрикам качества (MAE, RMSE, MAPE). Случайный лес занял второе место с близкими показателями к лидеру. Среди классических моделей SARIMA устойчиво превосходит Хольта–Уинтерса за счет явного учета автокорреляционной структуры ряда.

Сформулированы практические рекомендации по выбору метода в зависимости от длины ряда, требования к интерпретируемости и наличия дополнительных признаков.

Литература

1. Канторович Г. Г. Анализ временных рядов // Экономический журнал ВШЭ. — 2002. — Т. 6, № 1. — С. 79–103.
2. Орлов А. И. Эконометрика: учебник для вузов. — 4-е изд. — М.: Экзамен, 2007. — 576 с.
3. Айвазян С. А. Прикладная статистика. Основы эконометрики. Том 2. — М.: ЮНИТИ-ДАНА, 2001. — 432 с.
4. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин). Курс лекций. — М.: МФТИ, 2011.

5. Чистяков С. П. Случайные леса: обзор // Труды Карельского научного центра РАН. — 2013. — № 1. — С. 117–136.
6. Трегуб А. В., Трегуб И. В. Методика построения модели ARIMA для прогнозирования динамики временных рядов // Лесной вестник. — 2011. — № 5. — С. 179–183.
7. Афанасьев В. Н. Анализ временных рядов и прогнозирование : учебник. Саратов : Ай Пи Ар Медиа, 2019. 306 с.
8. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 400 с.
9. Документация библиотеки scikit-learn. URL: <https://scikit-learn.org/stable/> (дата обращения: 10.04.2026).
10. Хайкин С. Нейронные сети: полный курс. — 2-е изд., испр. — М.: Издательский дом «Вильямс», 2008. — 1104 с.
11. Документация библиотеки LightGBM. URL: <https://lightgbm.readthedocs.io/> (дата обращения: 10.04.2026).
12. Retailrocket Recommender System Dataset. Kaggle. — URL: <https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset> (дата обращения: 15.04.2026).

References

1. Kantorovich G. G. Time series analysis // HSE Economic Journal. 2002. Vol. 6, no. 1. Pp. 79–103. (In Russ.)
2. Orlov A. I. Econometrics: a textbook for universities. 4th ed. Moscow: Ekzamen, 2007. 576 p. (In Russ.)
3. Ayvazyan S. A. Applied statistics. Fundamentals of econometrics. Vol. 2. Moscow: YUNITI-DANA, 2001. 432 p. (In Russ.)

4. Vorontsov K. V. Machine learning methods based on precedents (machine learning theory). Lecture course. Moscow: MIPT, 2011. (In Russ.)
5. Chistyakov S. P. Random forests: a review // Transactions of the Karelian Research Centre of the Russian Academy of Sciences. 2013. No. 1. Pp. 117–136. (In Russ.)
6. Tregub A. V., Tregub I. V. Methodology for building an ARIMA model for forecasting time series dynamics // Lesnoy Vestnik / Forestry Bulletin. 2011. No. 5. Pp. 179–183. (In Russ.)
7. Afanasyev V. N. Time series analysis and forecasting: a textbook. Saratov: IPR Media, 2019. 306 p. (In Russ.)
8. Flach P. Machine learning. The art and science of algorithms that make sense of data. Moscow: DMK Press, 2015. 400 p. (In Russ.)
9. scikit-learn: machine learning in Python [Electronic resource]. URL: <https://scikit-learn.org/stable/> (date of access: 10.04.2026).
10. Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Moscow: Williams, 2008. 1104 p. (In Russ.)
11. LightGBM documentation [Electronic resource]. URL: <https://lightgbm.readthedocs.io/> (date of access: 10.04.2026).
12. Retailrocket Recommender System Dataset [Electronic resource]. URL: <https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset> (date of access: 15.04.2026).