Вестник Пермского университета. Серия «Экономика». 2024. Т. 19, № 2. С. 145–163.
Perm University Herald. Economy, 2024, vol. 19, no. 2, pp. 145–163.

УДК 338.27, ББК 65.050, JEL Code C1, C6, G17
DOI 10.17072/1994-9960-2024-2-145-163
EDN UBQMXB

Прогнозирование банковских продаж на примере ПАО «Сбербанк»

Анастасия Романовна Ермакова

Галина Сергеевна Васёва
РИНЦ Author ID: 775591, ✉ vasyova@econ.psu.ru

Пермский государственный национальный исследовательский университет, Пермь, Россия

Аннотация
Введение. В исследовании подчеркивается актуальность задачи моделирования и прогнозирования банковских продаж на примере ПАО «Сбербанк» в контексте эффективного управления бизнесом. Прогнозирование объемов продаж является важным инструментом, позволяющим предсказать спрос на продукты и услуги, определить оптимальные стратегии и тактики для достижения целей компании. Уникальность исследования состоит в использовании методов искусственного интеллекта в области маркетинга. Результаты применения методов прогнозирования на проприетарной выборке данных о ежедневных продажах ПАО «Сбербанк» обладают элементами новизны, что придает значимость разработке оптимальных стратегий и тактик для успешного управления бизнесом. Основная гипотеза исследования заключается в проверке прогностических способностей методов машинного обучения в сравнении с классическими эконометрическими подходами при моделировании объемов продажах ПАО «Сбербанк». *Цель.* Разработка моделей прогнозирования продаж универсальных продуктов и их инструментальная реализация для блока «Сеть продаж» ПАО «Сбербанк». *Материалы и методы.* В работе использованы методы системного анализа, статистические и экономико-математические методы анализа данных и их обработки. На собранных и предварительно обработанных данных о продажах условных продуктов ПАО «Сбербанк», отражающих динамику банковских продаж, проведены вычислительные эксперименты для построения ряда моделей прогнозирования и обоснован выбор наилучшей модели из числа построенных. *Результаты.* Модели на основе методов случайного леса (*Random Forest*) и градиентного бустинга (*XGBRegressor*) позволили получить прогнозы, точность которых существенно выше точности прогнозов *ARIMA*-модели и линейной регрессии на обучающей и тестовой выборках. *Выводы.* Результаты проведенной работы позволяют утверждать, что методы машинного обучения в настоящий момент являются перспективными для решения задач прогнозирования банковских продаж и могут выступать предметом дальнейших исследований в данной области. Внедрение методов машинного обучения в банковскую практику способно значительно улучшить эффективность существующего управления продажами и рисками.

Ключевые слова
Прогнозирование, объем продаж, финансовая отчетность, эконометрические модели, машинное обучение, статистические методы

Для цитирования
Ермакова А. Р., Васёва Г. С. Прогнозирование банковских продаж на примере ПАО «Сбербанк» // Вестник Пермского университета. Серия «Экономика». 2024. Т. 19, № 2. С. 145–163. DOI 10.17072/1994-9960-2024-2-145-163. EDN UBQMXB.

Конфликт интересов
Авторы заявляют об отсутствии конфликта интересов.

Статья поступила: 23.04.2024
Принята к печати: 25.05.2024
Опубликована: 01.07.2024

Forecasting of bank sales with Sberbank as a case study

Anastasia R. Ermakova

Galina S. Vasyova
RISC Author ID: 775591, ✉ vasyova@econ.psu.ru
Perm State University, Perm, Russia

Abstract
Introduction. This scientific study highlights the relevance of modeling and forecasting sales of Sberbank in terms of effective business management. Sales forecast is an important tool for predicting the demand for goods and services, as well as determining the adequate strategies and tactics to achieve the company's goals. The research is distinguished by its reference to artificial intelligence methods in the field of marketing. Forecasting methods applied to a proprietary data sample of Sberbank's daily sales give novel results, which reliably supports the development of adequate strategies and tactics for successful business management. The key hypothesis of the study is to check the prognostic potential of machine learning methods against the traditional econometric approaches to modeling Sberbank's sales. The *purpose* of the study is to develop sales forecasting models for multifunctional products and their practical instruments for Sberbank's Sales Network Block. *Materials and Methods*. The study relies on the methods of system-oriented analysis, statistical and economic mathematical methods of data analysis and their processing. Collected and pre-processed sales data for Sberbank's phantom products reflecting the dynamics of bank sales were used for computational experiments to build a few forecasting models and justify the choice of the best model among those built. *Results.* Random Forest and Gradient Busting (XGBRegressor) Models used training and test samples to give the forecasts with the accuracy significantly higher than the accuracy of forecasts by ARIMA-model and linear regression. *Conclusions.* The results of the analysis reliably confirm that machine learning methods are currently promising methods for forecasting bank sales and can be the subject of further research in this area. Machine learning techniques introduced into banking practices have the potential to significantly improve the effectiveness of existing sales and risk management.

Keywords
Forecasting, sales volume, financial reporting, econometric models, machine learning, statistical methods

For citation
Ermakova A. R., Vasyova G. S. Forecasting of bank sales with Sberbank as a case study. *Perm University Herald. Economy*, 2024, vol. 19, no. 2, pp. 145–163. DOI 10.17072/1994-9960-2024-2-145-163. EDN UBQMXB.

Declaration of conflict of interest: none declared.

Received: April 23, 2024
Accepted: May 25, 2024
Published: July 01, 2024

АНАЛИЗ БАНКОВСКИХ ПРОДАЖ

В современной рыночной экономике объем продаж является ключевым показателем эффективности деятельности в сфере материального производства и услуг. Вопросы организации продаж банковских продуктов становятся все более актуальными в связи с повышением требований клиентов к качеству обслуживания и усилением конкуренции. Продажи являются своеобразным индикатором эффективности всей деятельности банка [1–3].

Система продаж банковских продуктов не может существовать независимо от других структурных элементов банка, являясь частью его организационно-экономического механизма. Для оценки ее эффективности можно использовать несколько ключевых показателей [4; 5]:

– объем продаж – отражает количество и стоимость реализованных банковских продуктов за определенный период;

– конверсия – показывает процент успешных продаж от общего числа контактов с потенциальными клиентами;

– средний чек – позволяет определить среднюю стоимость продукта;

– коэффициент удержания клиентов;

– уровень удовлетворенности клиентов.

Для оценки эффективности системы продаж банковских продуктов также можно использовать различные методы анализа данных, такие как *ABC*-анализ, *SWOT*-анализ, анализ *KPI* и др. [6; 7].

Перечисленные показатели помогают оценить финансовую производительность и успешность компании, выявить ее прибыльность, эффективность и устойчивость к окружающей бизнес-среде.

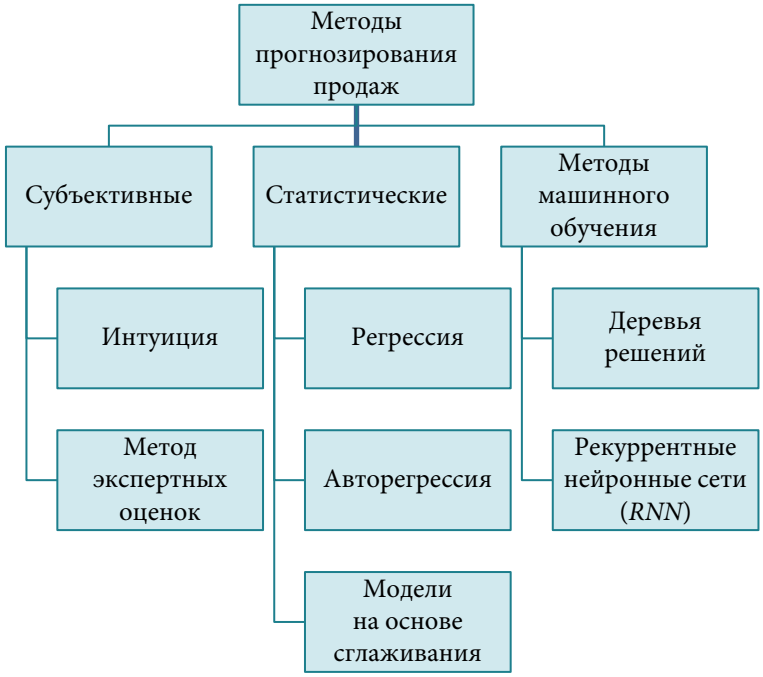
Предлагаем обратить внимание на один из ключевых показателей – объем продаж, который играет важную роль в оценке финансовой эффективности и успеха предприятия [4]. В ПАО «Сбербанк» данный показатель является основным в блоке «Сеть продаж» и измеряется в условных продуктах (далее – УП) или суммарных условных продуктах (далее – СУП). Каждый продукт, такой как кредит, депозит, кредитная карта, имеет уникальный вес в рамках УП. При этом вес УП может варьироваться в зависимости от конкретного продукта, подразделения и отдела. Это означает, что банк контролирует продажу определенных продуктов, увеличивая или уменьшая их вес в составе УП в соответствии с долгосрочными целями и стратегией. Такой подход помогает оптимизировать стратегии продаж и достигать наилучших результатов. Отметим, что в ходе исследования нами также рассмотрены понятие и сущность процесса продаж и система продаж банковских продуктов.

МЕТОДЫ ПРОГНОЗИРОВАНИЯ ПРОДАЖ

Прогнозирование продаж является важным инструментом для банковского сектора, позволяющим оптимизировать стратегию продаж, управлять запасами и планировать бюджет. Методы прогнозирования продаж можно разделить на субъективные, статистические и методы машинного обучения (рис. 1) [8; 9]. Каждая из этих групп имеет свои преимущества и недостатки, и нашей целью является поиск оптимальной модели прогнозирования продаж для ПАО «Сбербанк».

ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

Исследование предполагает прогнозирование объемов продаж ПАО «Сбербанк», данные о продажах предоставлены банком: они измеряются в СУП и охватывают период с 01 января 2020 г. по 07 апреля 2024 г., показанный в недельных временных интервалах в разрезе территориальных банков. Для исследования выбран язык *Python* – инструмент для анализа и прогнозирования временных рядов, использующий обширные библиотеки: *Pandas*, *NumPy*, *SciPy*, *XGBoost*, *statsmodels* и *scikit-learn*. Библиотеки *XGBoost* и *scikit-learn* предоставляют реализации алгоритмов градиентного бустинга



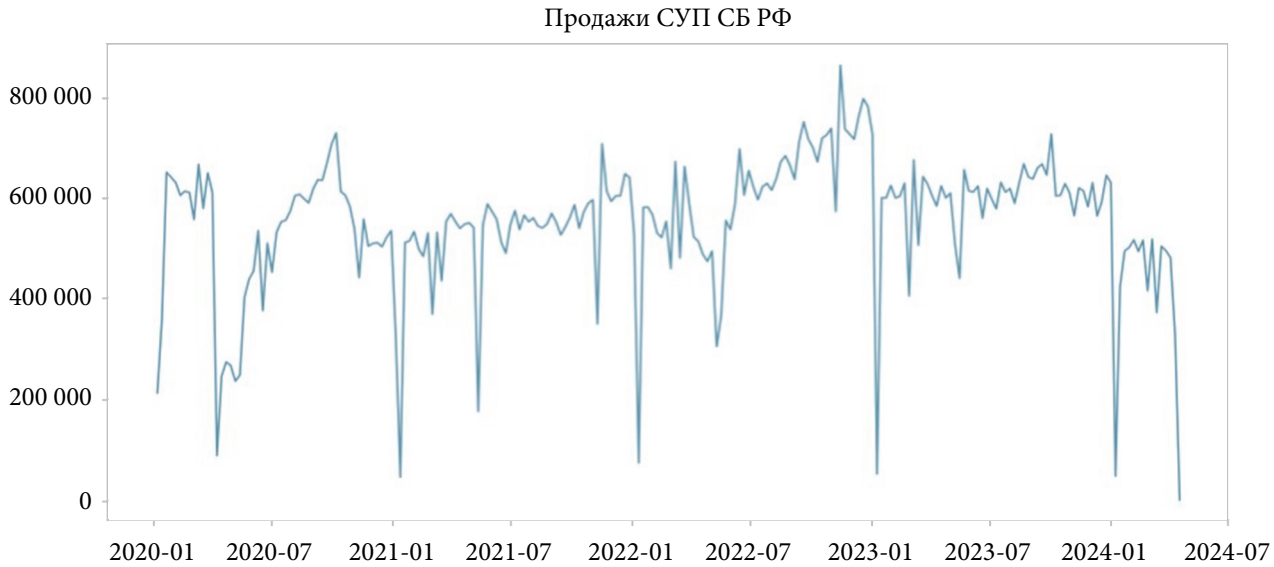
Источник: составлено авторами.

Рис. 1. Классификация методов прогнозирования

Fig. 1. Classification of forecasting methods

(*XGBoost*) и случайного леса (*Random Forest*), позволяя исследователям эффективно строить модели и делать прогнозы на основе временных данных. Эти инструменты обеспечивают высокую скорость обучения, хорошую обобщающую способность и возможность настройки гиперпараметров для достижения оптимальных результатов в анализе временных рядов.

Для первичного анализа рассмотрим агрегированное значение продаж ПАО «Сбербанк» по стране в целом (рис. 2). Наблюдается явная сезонность, выраженные праздничные дни, в то время как определенного тренда нет.



Источник: составлено авторами.

Рис. 2. Продажи ПАО «Сбербанк» за 2020–2024 гг.

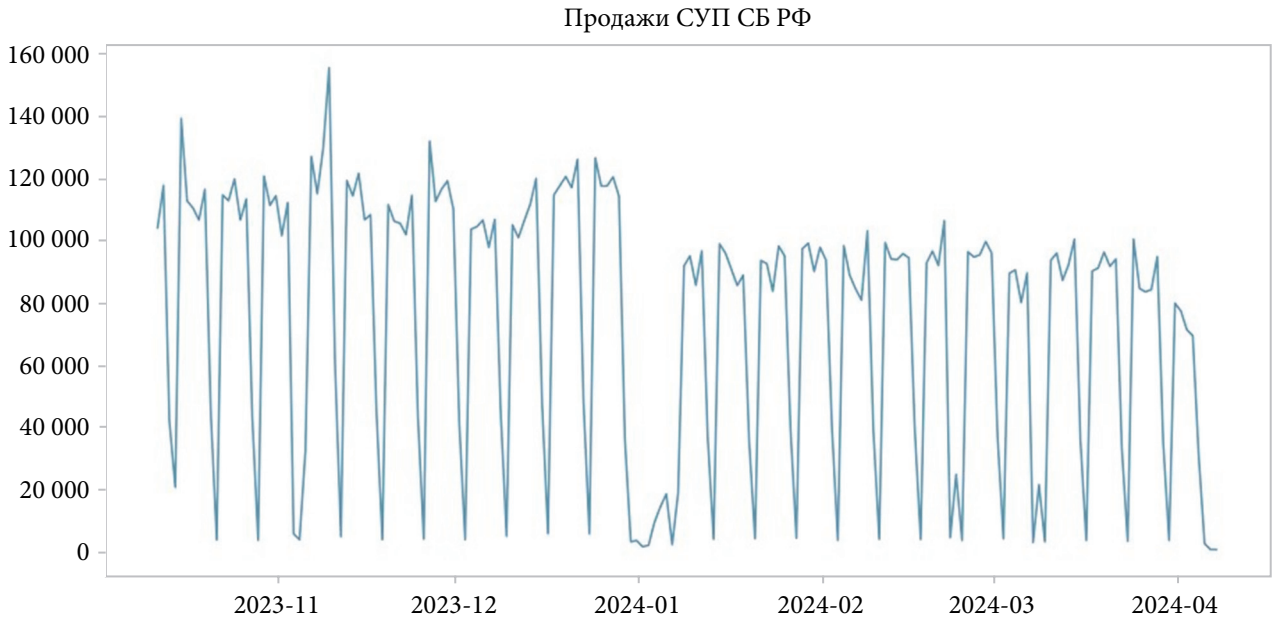
Fig. 2. Sberbank's sales, 2020–2024

Перед нами стоит задача краткосрочного дневного прогноза, поэтому рассмотрим данные за последние шесть месяцев: с 11 октября 2023 г. до 07 апреля 2024 г. (рис. 3).

Наблюдается еженедельная сезонность, а также низкие объемы продаж во время новогодних праздников. Отметим, что с начала нового года произошло изменение методологии и пересчет СУП банка.

Проведенный на следующем шаге исследования анализ пропущенных данных позволил установить, что пропуски в них отсутствуют.

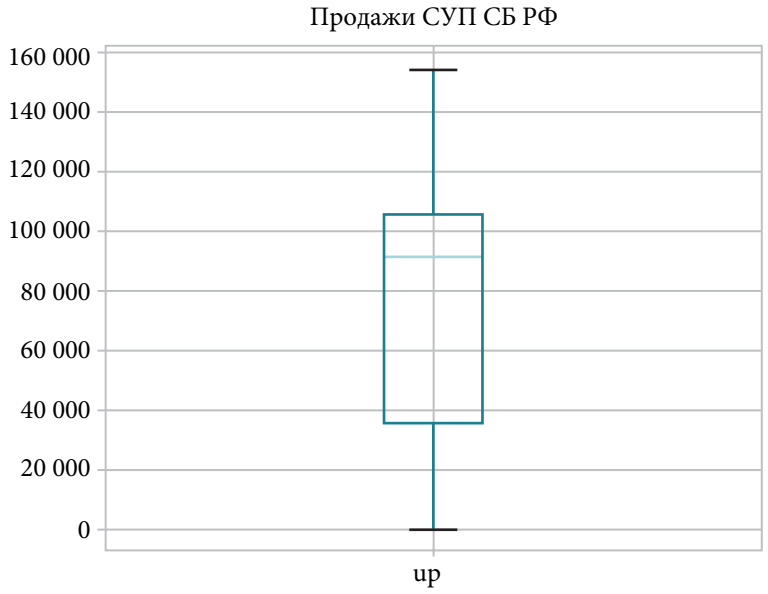
Далее нами оценены данные на выбросы (рис. 4), которых также не обнаружено.



Источник: составлено авторами.

Рис. 3. Продажи ПАО «Сбербанк» в период 11.10.2023–07.04.2024

Fig. 3. Sberbank's sales from 11.10.2023 to 07.04.2024



Источник: составлено авторами.

Рис. 4. «Ящик с усами» продаж ПАО «Сбербанк»

Fig. 4. Sberbank's sales boxplot

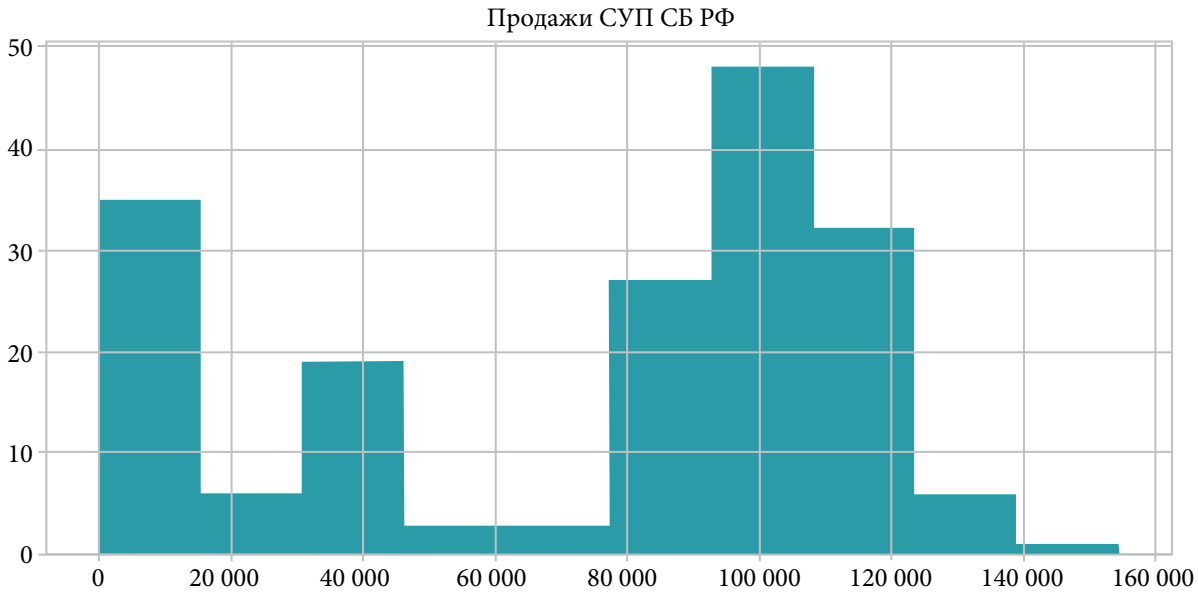
Произведем анализ описательных статистик на основе данных за предшествующие полгода относительно объемов продаж в ПАО «Сбербанк». В ходе анализа установлено, что средний объем продаж составляет 72 701,56 СУП, а диапазон значений колеблется в интервале между 0,28 и 154 918,66 СУП. Результаты представлены в табл. 1.

Изучим также распределение данных с помощью гистограммы, которая нужна для более глубокого понимания их структуры и выявления возможных закономерностей или аномалий. Исходя из анализа рис. 5, мы видим, что в нашем случае данные не имеют нормального распределения.

Таким образом, в качестве зависимой (экзогенной) переменной нами выбраны продажи, а в качестве независимых (эндогенных) – данные, описывающие текущий день (табл. 2).

ЭКОНОМЕТРИЧЕСКИЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Линейная регрессия – это статистический метод, который используется для определения связи между зависимой переменной (прогнозируемой) и одной или несколькими независимыми переменными. Иначе говоря, суть данного метода заключается в стремлении найти



Источник: составлено авторами.

Рис. 5. Гистограмма продаж ПАО «Сбербанк»

Fig. 5. Sberbank's sales histogram

Табл. 1. Описательные статистики продаж

Table 1. Descriptive statistics of sales

Показатель	Значение
Среднее значение	72 701,5613
Стандартное отклонение	43 080,2920
Минимальное значение	0,2800
Максимальное значение	154 918,6600
Мода	91 757,4750
Квантиль уровня 25%	35 380,1850
Квантиль уровня 75%	105 835,0050

Источник: составлено авторами.

Табл. 2. Обозначение показателей

Table 2. Description of indicators

Показатель	Тип показателя	Ед. изм.	Обозначение
Продажи ПАО «Сбербанк»	Количественный	СУП	Y
День недели	Качественный	1–7	X_1
День месяца	Качественный	1–31	X_2
Праздничные дни	Качественный	0 / 1	X_3

Источник: составлено авторами.

линейную зависимость между зависимой переменной и одной или несколькими независимыми переменными [10].

Уравнение в общем виде выглядит как

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_p X_{pt} + \varepsilon_t, \quad (1)$$

где Y_t – значение временного ряда в момент времени t; $X_{1t}, X_{2t}, \ldots, X_{pt}$ – независимые переменные в моменты времени t; $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ – коэффициенты регрессии; ε_t – случайная ошибка в момент времени t.

Цель линейной регрессии для временных рядов состоит в оценке коэффициентов $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ таким образом, чтобы минимизировать сумму квадратов ошибок ε_t.

Уравнение модели (2) показывает, как зависимая переменная Y_t связана с независимыми переменными X_{1t}, X_{2t} и X_{3t} с определенными коэффициентами:

$$Y_t = 122170{,}47 - 51109{,}37 X_{1t} - \\ - 14388{,}03 X_{2t} + 253{,}45 X_{3t}. \quad (2)$$

Результаты построения модели отражены на рис. 6.

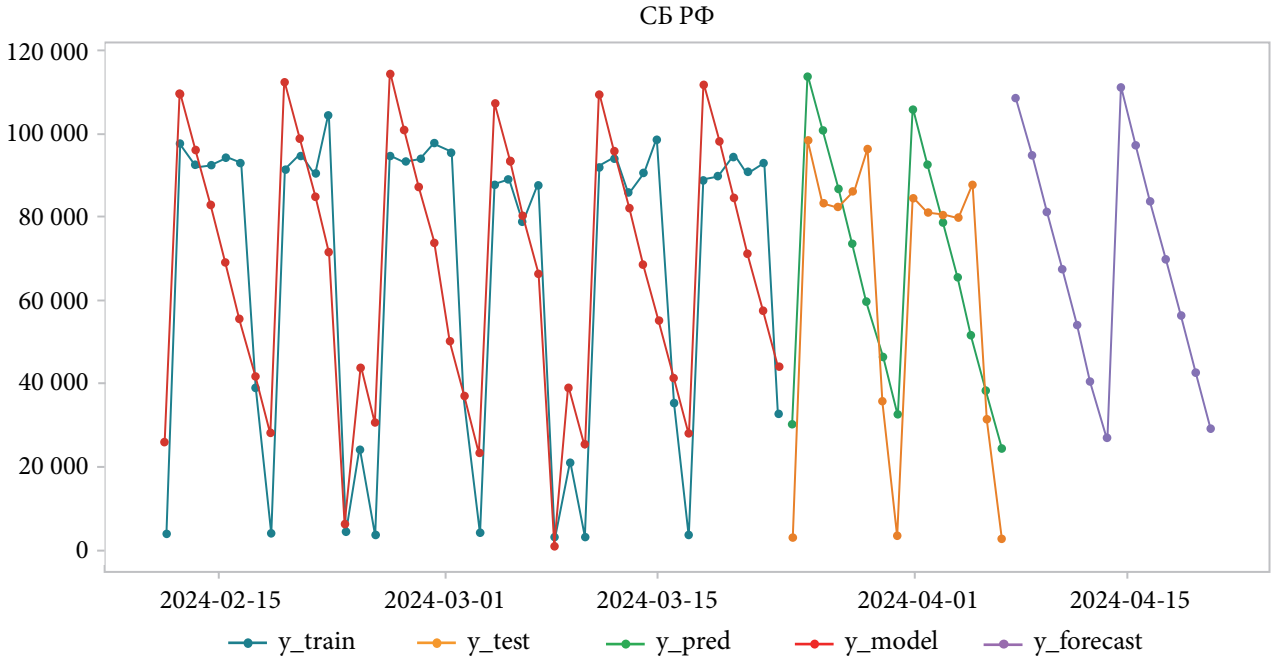
После построения модели проведена работа по оценке качества, результаты которой представлены в табл. 3.

Полученные данные говорят о том, что модель линейной регрессии демонстрирует удовлетворительные результаты при прогнозировании зависимой переменной.

Далее перейдем к рассмотрению авторегрессионной модели интегрированного скользящего среднего. Рассмотрим три компонента, которые интегрирует в себе *ARIMA*-модель (*Autoregressive Integrated Moving Average*)[1].

1. *AR* (авторегрессионный термин) – относится к использованию прошлых значений временного ряда для прогнозирования будущих значений. Параметр p в модели авторегрессии определяет количество прошлых значений, которые используются для прогнозирования:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t. \quad (3)$$



Источник: составлено авторами.

Рис. 6. Модель линейной регрессии

Fig. 6. A linear regression model

[1] *Магнус Я. Р., Катышев П. К., Пересецкий А. А.* Эконометрика. Начальный курс: учебник. 6-е изд., перераб. и доп. М.: Дело, 2004. 576 с.

Табл. 3. Оценки качества линейной регрессии

Table 3. Accuracy of linear regression

Название	Обозначение	Формула	Значение
Коэффициент детерминации	R^2	$1-\frac{\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$	0,59
Средняя квадратичная ошибка	*MSE*	$\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2$	432 423 051,80
Средняя абсолютная ошибка	*MAE*	$\frac{1}{N}\sum_{i=1}^{n}\lvert a(x_i)-y_i\rvert$	17 290,81
Корень из средней квадратичной ошибки	*RMSE*	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}$	20 794,78
Средняя абсолютная процентная ошибка	*MAPE*	$\frac{1}{N}\sum_{i=1}^{n}\frac{\lvert a(x_i)-y_i\rvert}{\lvert y_i\rvert}\times 100\%$	156,25
Взвешенная абсолютная процентная ошибка	*WAPE*	$\frac{\sum_{i=1}^{n}\lvert Y_i-e_i\rvert}{\sum_{i=1}^{n}\lvert Y_i\rvert}\times 100\%$	26,76

Источник: составлено авторами.

Параметр можно определить по *PACF* (*Partial Auto-Correlation Function*) – «частной корреляционной функции» между Y_t и Y_{t-k} при исключении влияния $Y_{t-1}, \ldots, Y_{t-k+1}$.

2. *MA* (скользящее среднее) – используется для учета прошлых ошибок прогнозов и их влияния на будущие значения. Параметр q определяется по автокорреляционной функции (*Auto-Correlation Function, ACF*):

$$\rho_k = \frac{cov\{Y_t, Y_{t-k}\}_t}{var\{Y_t\}_t}, \quad (4)$$

$$Y_t = \varepsilon_t + \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2} + \ldots + \alpha_q\varepsilon_{t-q}, \quad (5)$$

где ε_t – белый шум, всегда являющийся стационарным процессом.

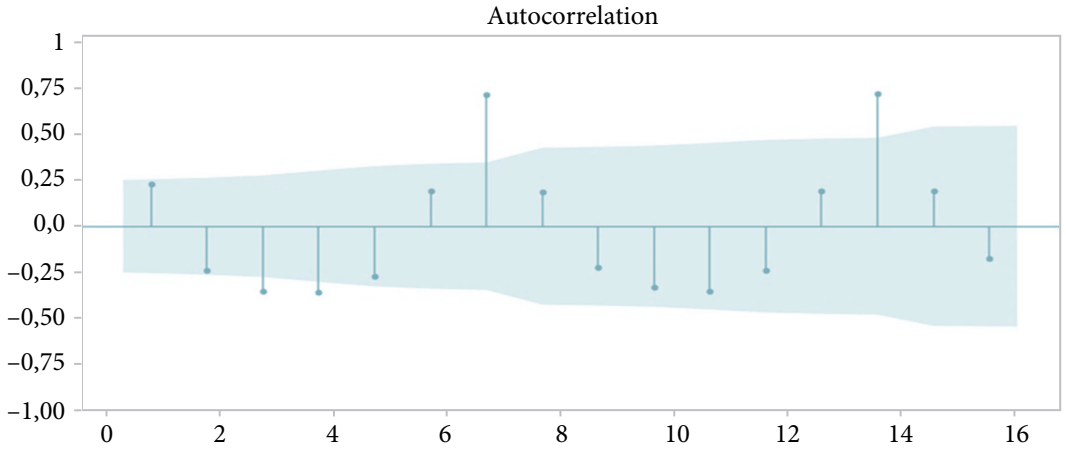
Скользящее среднее показывает наличие колебаний в ряду. Чем выше значение скользящего среднего, тем выше вероятность колебаний.

3. *I* (интегрирующий член) – используется для работы с нестационарными данными. Если временной ряд не является стационарным, применяется операция разности порядка d для его преобразования в стационарный ряд. Параметр d определяется с помощью тестов, таких как *ADF* и *KPSS*, которые позволяют определить степень дифференцирования, необходимую для стационарности.

Будем следовать методологии Бокса–Дженкинса для подбора оптимальной *ARIMA*-модели [11; 12]. Начнем с построения *ACF* и *PACF* (рис. 7–8 соответственно). По автокорреляционной функции на рис. 7 и 8 видно, что присутствует сезонность.

Предположим, что коэффициенты p и q для модели будут равны 7 и 7. Затем проведем анализ временного ряда путем его декомпозиции на тренд, сезонную составляющую и остатки. Для этого используем аддитивную и мультипликативную модели, а также *LOESS*-модель с целью декомпозиции. После оценки остатков выявлено, что наилучшим вариантом является мультипликативная модель. Результаты декомпозиции представлены на рис. 9, где четко прослеживаются сезонность и тренд, связанный с праздничными днями.

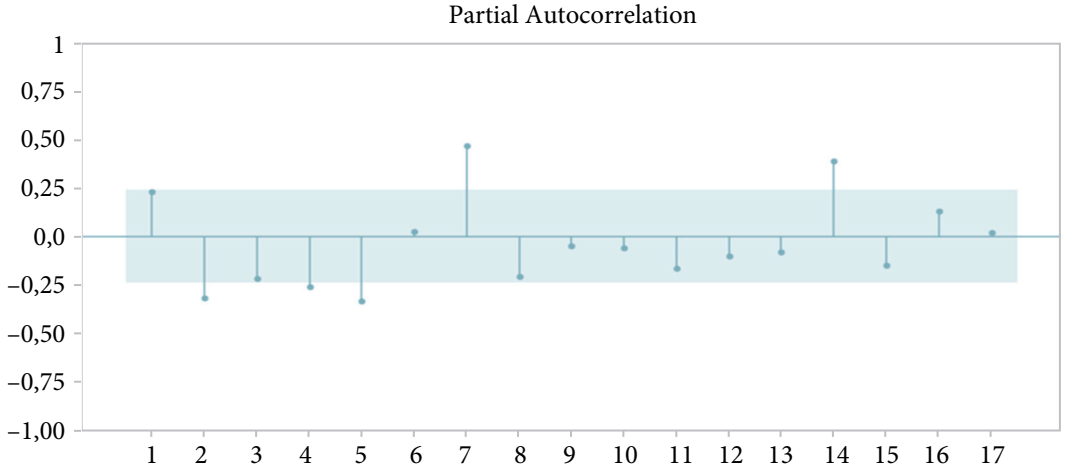
Далее изучим тренд, построив линейную, квадратичную, кубическую и экспоненциальную модели (рис. 10).



Источник: составлено авторами.

Рис. 7. Автокорреляционная функция временного ряда

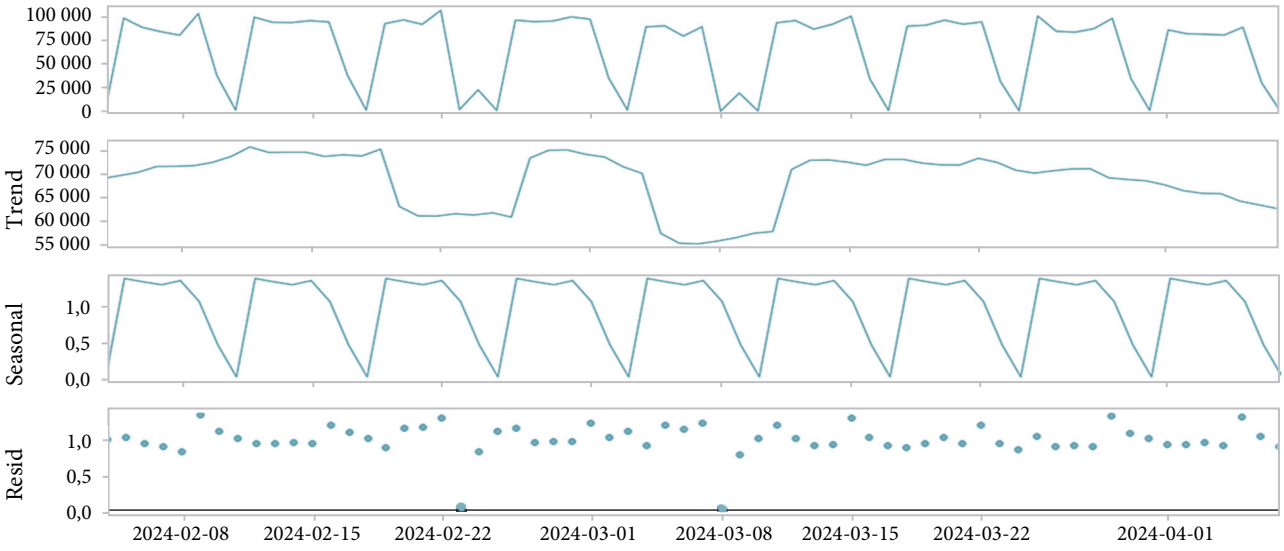
Fig. 7. Autocorrelation function of time series



Источник: составлено авторами.

Рис. 8. Частная автокорреляционная функция временного ряда

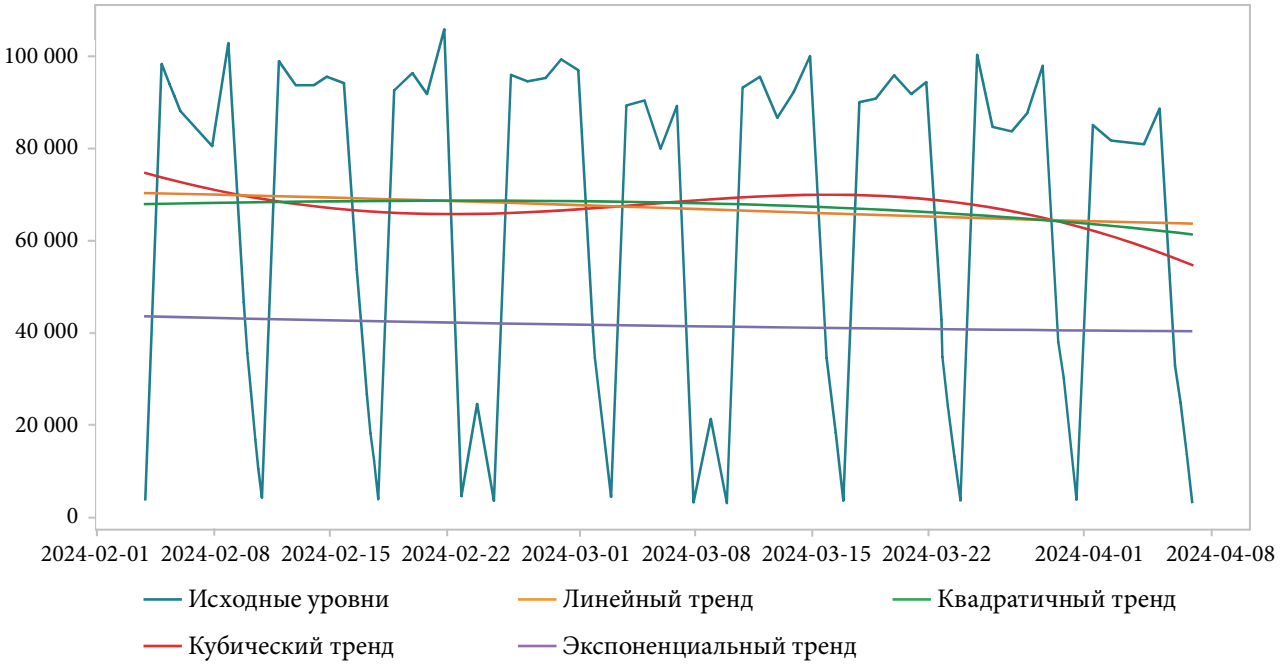
Fig. 8. Partial autocorrelation function of time series



Источник: составлено авторами.

Рис. 9. Декомпозиция временного ряда

Fig. 9. Time series decomposition



Источник: составлено авторами.

Рис. 10. Тренд временного ряда
Fig. 10. Time series trend

Можно отметить, что модели на рис. 10 показывают отсутствие ярко выраженного тренда, поскольку в данных не прослеживается четкое направление роста (спада) во времени.

Далее проверим ряд на стационарность: сезонность ряда видна визуально, ряд нестационарен. В связи с этим принимаем решение использовать *SARIMA*-модель вместо *ARIMA* из-за наличия сезонности. После тщательного подбора параметров и тестирования различных моделей определяем, что оптимальной является модель *SARIMAX* (7, 1, 7), уравнение которой выглядит следующим образом:

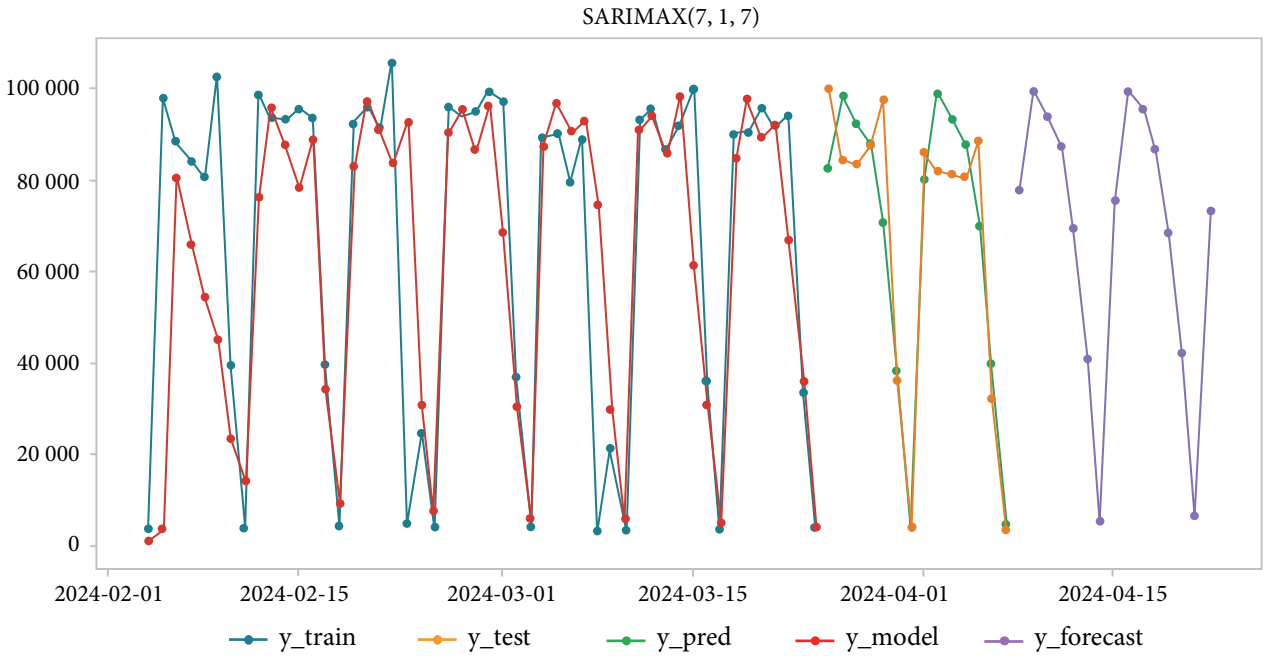
$$\begin{aligned}\Delta_t = {} & 3{,}93\times10^{8} - 0{,}87\Delta_{t-1} - 0{,}86\Delta_{t-2} - \\ & - 0{,}88\Delta_{t-3} - 0{,}86\Delta_{t-4} - 0{,}88\Delta_{t-5} - \\ & - 0{,}86\Delta_{t-6} + 0{,}12\Delta_{t-7} + \varepsilon_t - 0{,}07\varepsilon_{t-1} - \\ & - 0{,}03\varepsilon_{t-2} + 0{,}07\varepsilon_{t-3} - 0{,}08\varepsilon_{t-4} + \\ & + 0{,}07\varepsilon_{t-5} - 0{,}03\varepsilon_{t-6} - 0{,}93\varepsilon_{t-7},\end{aligned} \qquad (6)$$

где Δ_t – разность между значениями временного ряда на момент времени t и $t - 1$.

В модели присутствует 7 лагов, что соответствует недельной сезонности. Это означает, что мы учитываем информацию о данных за предыдущие 7 дней при прогнозировании текущего значения. Использование такой сезонной составляющей полезно для улавливания цикличности или паттернов, повторяющихся еженедельно во временном ряду. Результаты модели (6) представлены на рис. 11.

Затем модель была оценена. Результаты оценки, показанные в табл. 4, позволяют сделать вывод, что рассматриваемая модель в целом адекватно описывает данные, хотя и наблюдается некоторое расхождение между прогнозами и фактическими значениями. Такой вывод основан на значении средней абсолютной ошибки (*MAE*), равном 10 041,65, поскольку *MAE* является мерой разницы между фактическими и прогнозируемыми значениями временного ряда (чем ближе этот показатель к нулю, тем лучше прогнозные качества модели).

Таким образом, нами построены две модели с использованием статистических методов прогнозирования временных рядов, обе оценены как удовлетворительные. На следующем шаге перейдем к методам машинного обучения, использование которых позволит более гибко учитывать сложные взаимосвязи данных и улучшить качество прогнозов.



Источник: составлено авторами.

Рис. 11. Модель *SARIMAX* (7, 1, 7)

Fig. 11. SARIMAX Model (7, 1, 7)

Табл. 4. Оценки качества SARIMAX (7, 1, 7)

Table 4. SARIMAX scores (7, 1, 7)

Название	Обозначение	Формула	Значение
Средняя квадратичная ошибка	*MSE*	$\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2$	161 752 430,30
Средняя абсолютная ошибка	*MAE*	$\frac{1}{N}\sum_{i=1}^{n}\lvert a(x_i)-y_i\rvert$	10 041,65
Корень из средней квадратичной ошибки	*RMSE*	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}$	12 718,19
Средняя абсолютная процентная ошибка	*MAPE*	$\frac{1}{N}\sum_{i=1}^{n}\frac{\lvert a(x_i)-y_i\rvert}{\lvert y_i\rvert}\times 100\%$	17,74
Взвешенная абсолютная процентная ошибка	*WAPE*	$\frac{\sum_{i=1}^{n}\lvert Y_i-e_i\rvert}{\sum_{i=1}^{n}\lvert Y_i\rvert}\times 100\%$	14,93

Источник: составлено авторами.

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Многие методы машинного обучения применяют деревья для решения задач классификации и регрессии. Решающие деревья представляют собой структуру в виде графа, где узлы содержат признаки для разделения выборки, а листья представляют собой части выборки. Глубина дерева определяется количеством уровней иерархии в структуре [12].

«Одно дерево – это хорошо, а много – еще лучше, а когда деревьев много – это уже лес»[1]. Объединение множества решающих деревьев дает композицию алгоритмов, одной из разновидностей которой является случайный лес. Объединение множества слабых алгоритмов с невысокой точностью дает один сильный алгоритм с хорошей точностью.

В методе случайного леса обучают каждый алгоритм из композиции, а ответом является усредненный результат по всем алгоритмам, входящим в композицию. В случае регрессии ответ $a(x)$ находится по формуле

$$a(x) = \frac{1}{N}\sum_{n=1}^{N} b_n(x), \qquad (7)$$

где $b_n(x)$ – предсказание n-го базового алгоритма на входных данных x [12].

Каждый алгоритм в композиции дает собственное предсказание, которое усредняется для получения итогового результата $a(x)$. Обучение деревьев происходит независимо друг от друга (на разных подмножествах).

Метод *Random Forest* обладает рядом преимуществ, включая высокую точность предсказаний за счет использования ансамбля деревьев, устойчивость к переобучению благодаря случайному выбору признаков и данных для построения каждого дерева, способность обрабатывать большое количество признаков без необходимости предварительной обработки данных [13; 14]. Однако у данного метода существуют ограничения, такие как склонность к переобучению при использовании большого количества деревьев или при наличии шумных данных, неинтерпретируемость результатов из-за большого количества деревьев и их комбинаций, а также вычислительная сложность при построении и обучении леса деревьев, особенно при работе с большими объемами данных. В целом *Random Forest* является мощным алгоритмом машинного обучения с высокой точностью и устойчивостью, но при его применении необходимо учитывать указанные ограничения и особенности для эффективного использования в конкретной задаче [14].

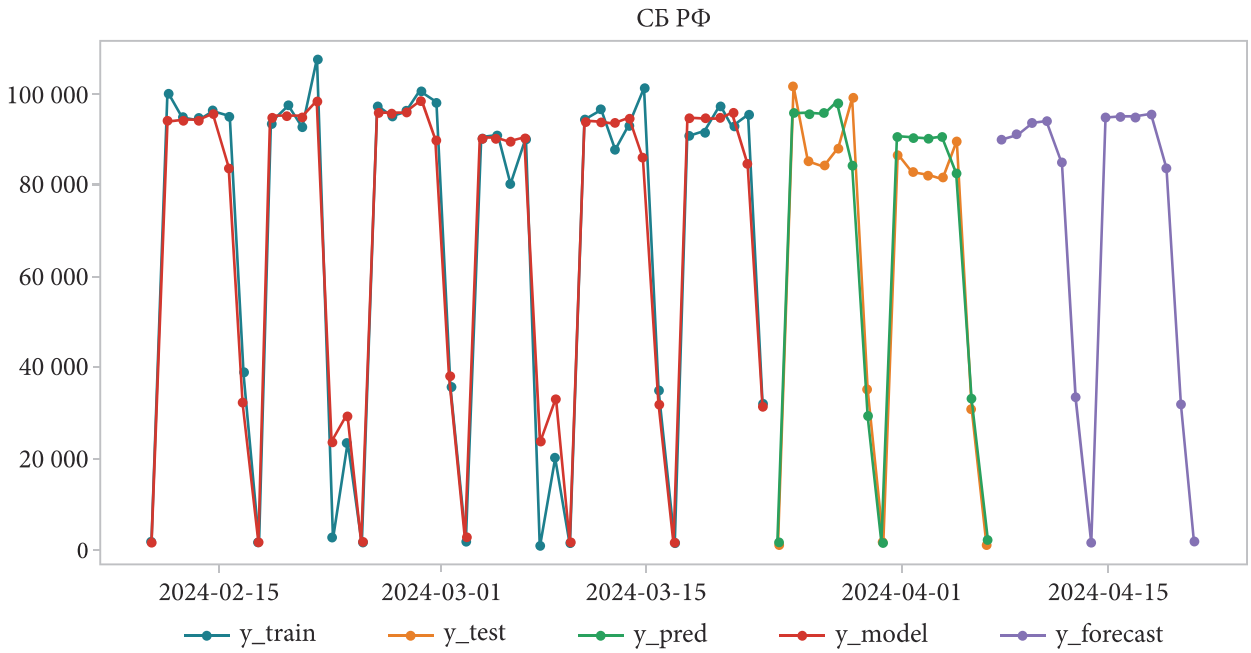
Благодаря использованию модели *Random Forest Regressor* нам удалось создать эффективную модель. Графически результаты модели представлены на рис. 12, а оценка ее качества – в табл. 5. Следует отметить, что средняя абсолютная ошибка уменьшилась практически вдвое – до 5 921,26 суммарных агрегированных продуктов. Этот показатель свидетельствует о том, что модель *Random Forest Regressor* лучше соответствует данным и способна делать более точные прогнозы по сравнению с предыдущими моделями.

Далее рассмотрим метод бустинга. Усиление (*boosting*) – это метод обучения, который строит композицию из базовых алгоритмов для повышения их эффективности. В отличие от бэггинга (от англ. *bootstrap aggregating*), где модели работают независимо, в бустинге модели приспосабливаются к данным последовательно, исправляя ошибки предыдущих моделей.

Градиентный бустинг использует соответственно алгоритм градиентного спуска для добавления новых слабых алгоритмов в композицию. При этом находится оптимальный вектор сдвига, который улучшает работу предыдущих алгоритмов. Он вычисляется как антиградиент функции ошибок предыдущей композиции[2]. Таким образом мы определяем, какие значения должны принимать объекты обучающей выборки для минимизации отклонения ответов от истинных значений при добавлении нового алгоритма в композицию [15].

[1] *Лимановская О. В., Алферьева Т. И.* Основы машинного обучения: учеб. пособие. Екатеринбург: Изд-во Урал. ун-та, 2020. 88 с.

[2] *Кугаевских А. В., Муромцев Д. И., Кирсанова О. В.* Классические методы машинного обучения. СПб.: Университет ИТМО, 2022. 53 с.



Источник: составлено авторами.

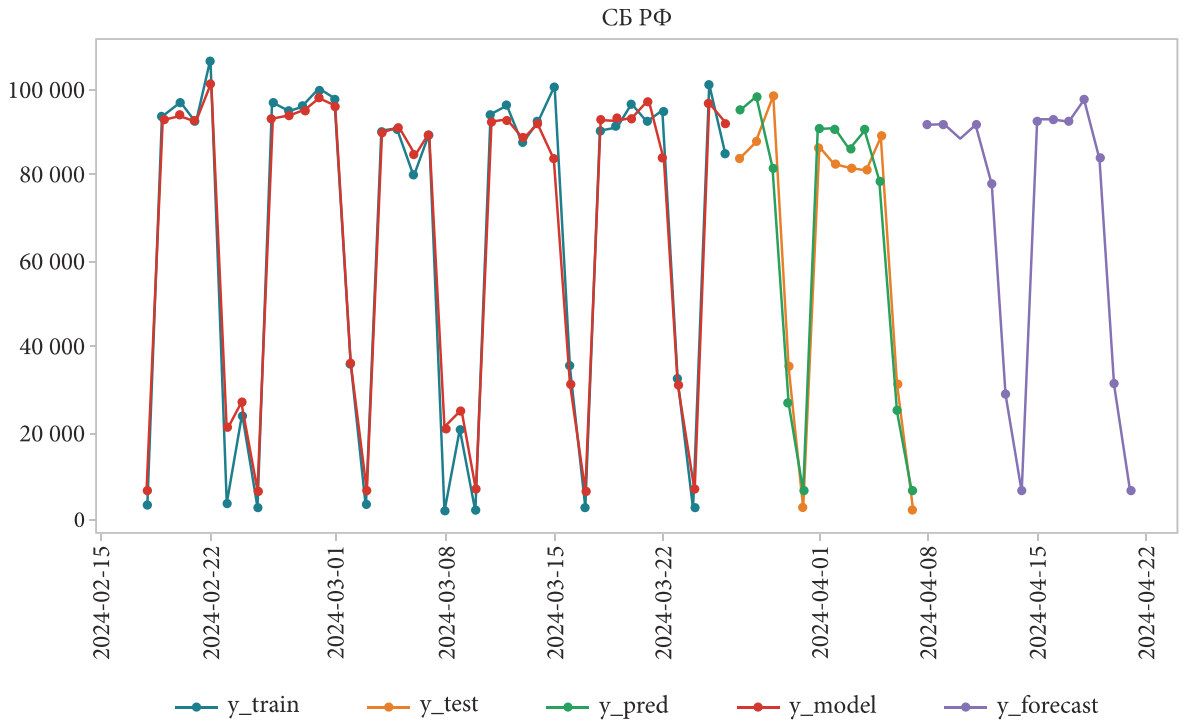
Рис. 12. Модель *Random Forest*
Fig. 12. Random Forest model

Табл. 5. Оценки качества модели *Random Forest*
Table 5. Random Forest Model's scores

Название	Обозначение	Формула	Значение
Коэффициент детерминации	R^2	$1-\frac{\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$	0,95
Средняя квадратичная ошибка	*MSE*	$\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2$	52 539 760,07
Средняя абсолютная ошибка	*MAE*	$\frac{1}{N}\sum_{i=1}^{n}\left\|a(x_i)-y_i\right\|$	5 921,26
Корень из средней квадратичной ошибки	*RMSE*	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}$	7 248,43
Средняя абсолютная процентная ошибка	*MAPE*	$\frac{1}{N}\sum_{i=1}^{n}\frac{\left\|a(x_i)-y_i\right\|}{\left\|y_i\right\|}\times 100\%$	18,28
Взвешенная абсолютная процентная ошибка	*WAPE*	$\frac{\sum_{i=1}^{n}\left\|Y_i-e_i\right\|}{\sum_{i=1}^{n}\left\|Y_i\right\|}\times 100\%$	9,16

Источник: составлено авторами.

Нами была разработана модель с применением *XGBRegressor*, которая оказалась успешной. На рис. 13 изображен график, отражающий результаты модели, а показатели ее качества представлены в табл. 6. Следует отметить, что средняя абсолютная ошибка составила 8 059,20 суммарных агрегированных продуктов. Это указывает на то, что модель с *XGBRegressor* является релевантной и хорошо адаптирована к данным.



Источник: составлено авторами.

Рис. 13. Модель на основе градиентного бустинга
Fig. 13. Model based on gradient boosting

Табл. 6. Оценки качества модели *XGBRegressor*
Table 6. XGBRegressor model's scores

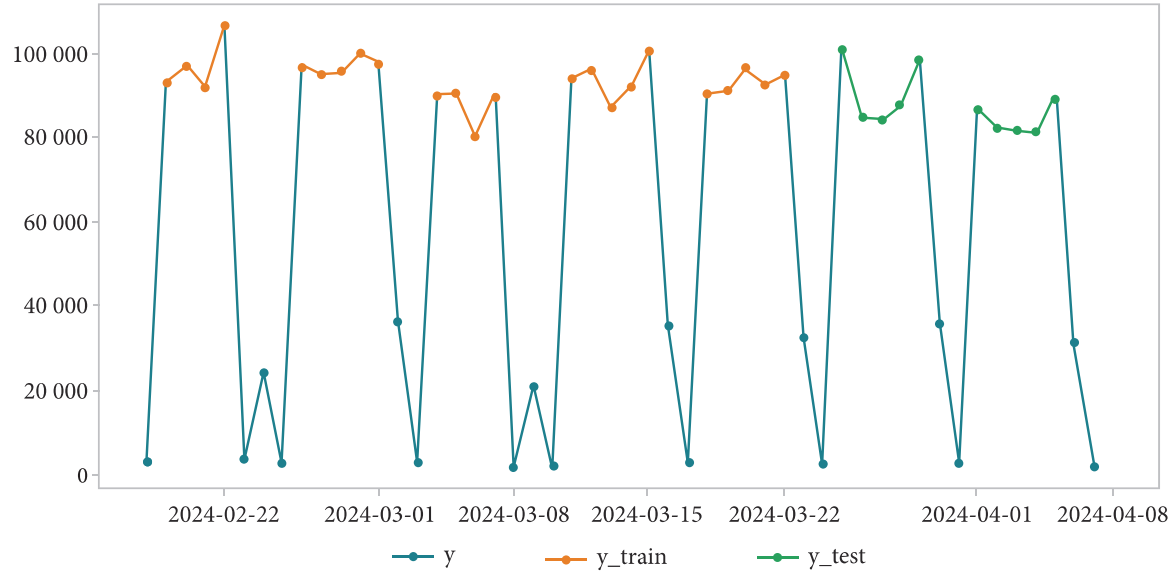
Название	Обозначение	Формула	Значение
Коэффициент детерминации	R^2	$1-\frac{\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$	0,93
Средняя квадратичная ошибка	*MSE*	$\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2$	36 565 554,92
Средняя абсолютная ошибка	*MAE*	$\frac{1}{N}\sum_{i=1}^{n}\lvert a(x_i)-y_i\rvert$	8 059,20
Корень из средней квадратичной ошибки	*RMSE*	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}$	8 813,48
Средняя абсолютная процентная ошибка	*MAPE*	$\frac{1}{N}\sum_{i=1}^{n}\frac{\lvert a(x_i)-y_i\rvert}{\lvert y_i\rvert}\times 100\%$	36,83
Взвешенная абсолютная процентная ошибка	*WAPE*	$\frac{\sum_{i=1}^{n}\lvert Y_i-e_i\rvert}{\sum_{i=1}^{n}\lvert Y_i\rvert}\times 100\%$	12,72

Источник: составлено авторами.

Итак, авторами были изучены методы машинного обучения, модели на основе которых продемонстрировали более высокие показатели качества, чем при использовании статистических методов. Несмотря на полученный результат, попытаемся усовершенствовать модели для достижения еще более точных и надежных результатов.

С этой целью рассмотрим идею декомпозиции данных на основе недельной сезонности. В результате проведенного анализа мы выяснили, что в полученном временном ряду присутствует недельная цикличность, поэтому разделим данные на три категории: будние дни, суббота и воскресенье. По нашим предположениям, такой подход позволит повысить точность модели, для построения которой выберем метод случайного леса (*Random Forest*), поскольку он продемонстрировал лучшие результаты на данном наборе данных. Таким образом, выбранный подход позволит учесть сезонные колебания и повысить качество прогнозов.

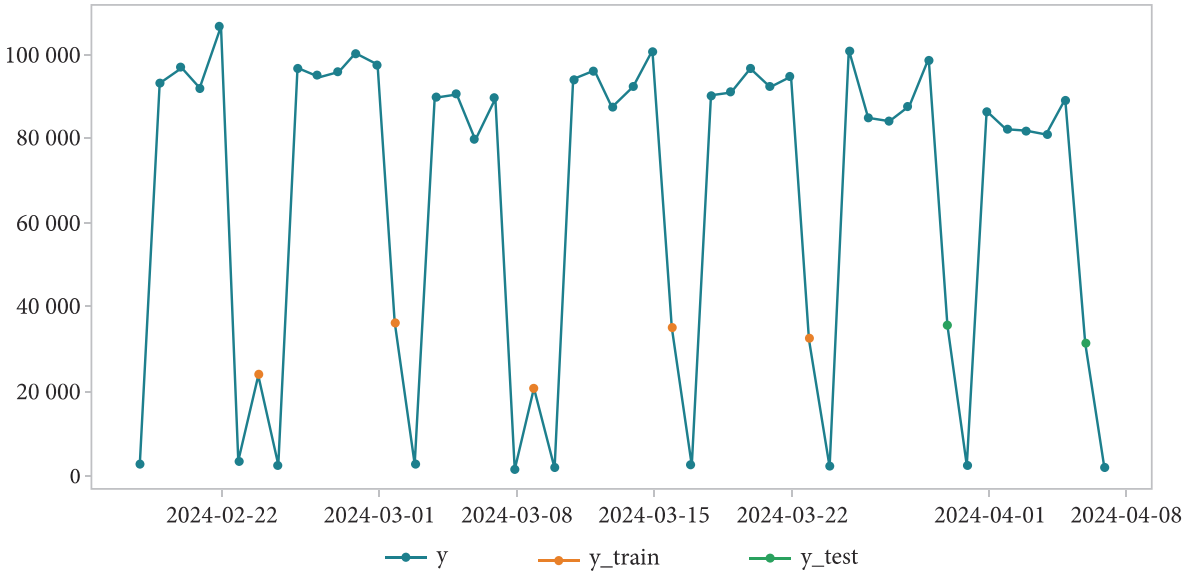
Мы разделили имеющиеся данные на три категории и выделили для каждой из них тестовую и обучающую выборки. Графически результаты этого разделения представлены на рис. 14–16. Благодаря используемому подходу мы сможем провести эффективное обучение моделей для каждой категории данных и оценить их качество на соответствующих тестовых выборках.



Источник: составлено авторами.

Рис. 14. Временной ряд будних дней

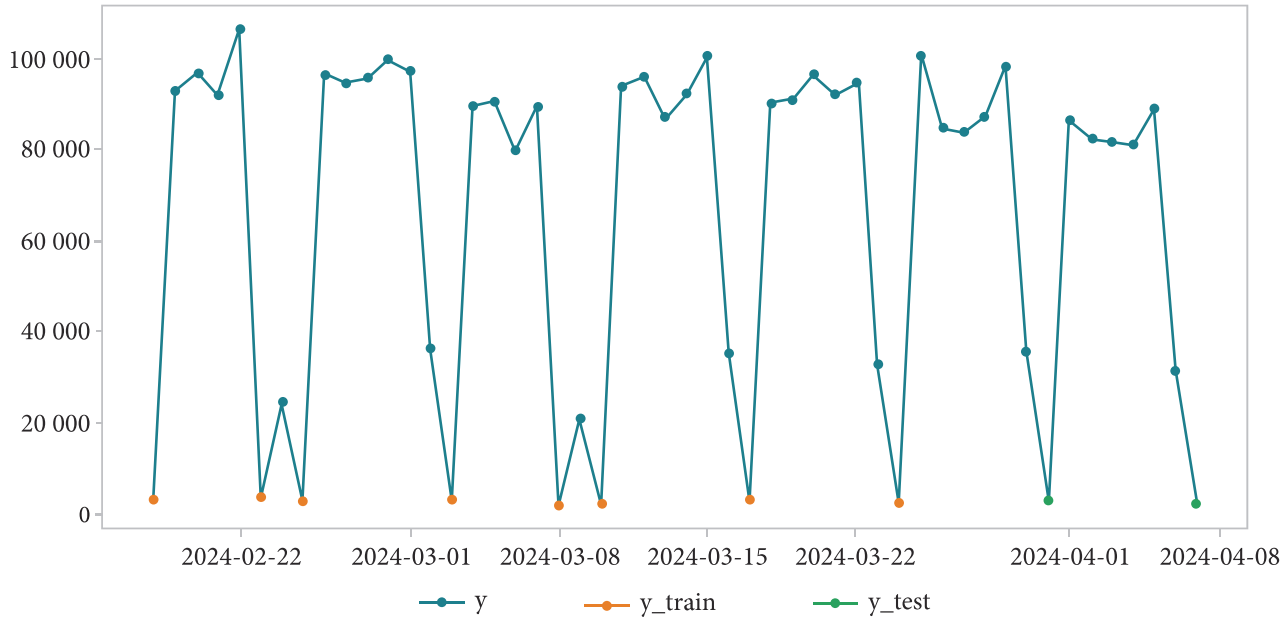
Fig. 14. Time series of weekdays



Источник: составлено авторами.

Рис. 15. Временной ряд субботних дней

Fig. 15. Time series of Saturdays



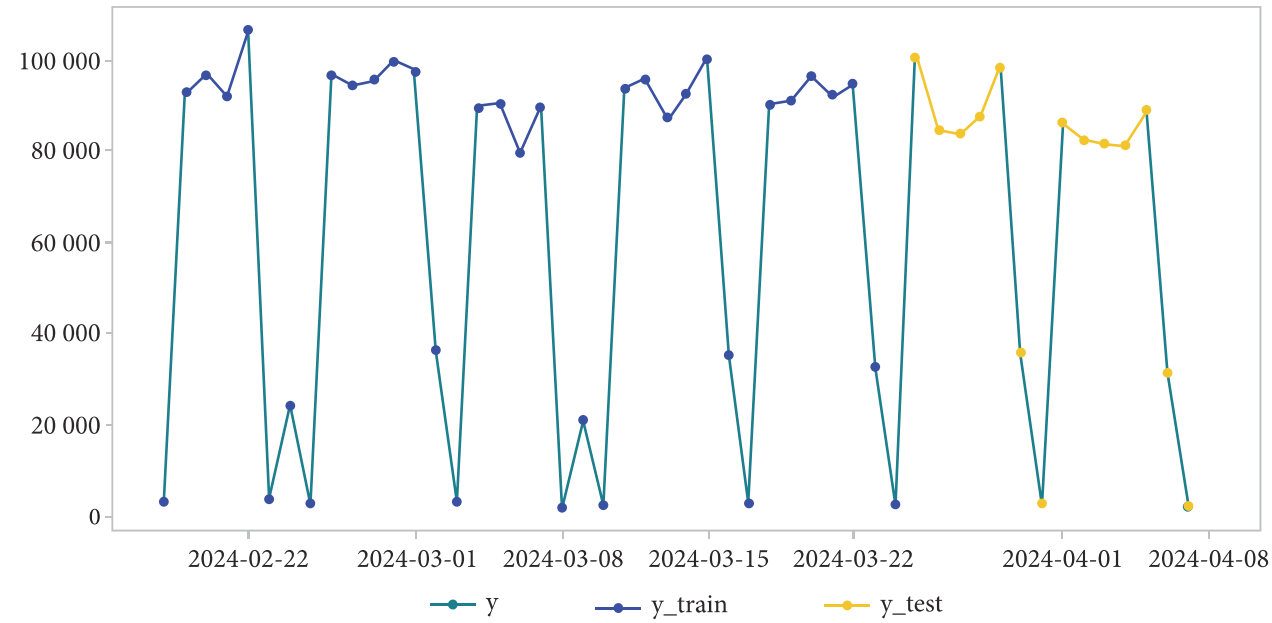
Источник: составлено авторами.

Рис. 16. Временной ряд воскресных дней

Fig. 16. Time series of Sundays

На рис. 17 представлена комплексная визуализация трех категорий.

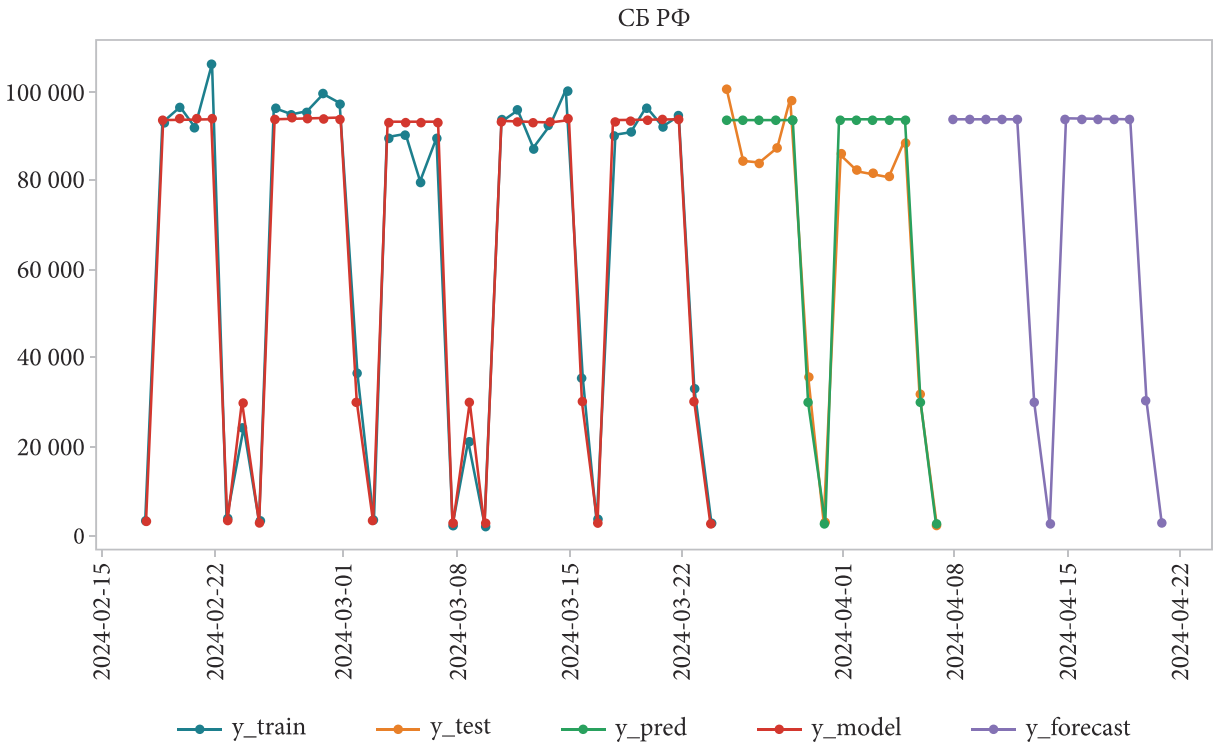
Далее построим модели для каждой категории данных сначала по отдельности, затем объединим их в общую модель. Каждая из обозначенных моделей построена с использованием метода случайного леса (рис. 18). Этот подход позволил учесть особенности и сезонность в каждой категории данных, а затем объединить их для получения общего прогностического результата. Результаты оценки качества данной модели представлены в табл. 7.



Источник: составлено авторами.

Рис. 17. Деление данных на обучающую и тестовую выборки

Fig. 17. Data grouped into training and test samples



Источник: составлено авторами.

Рис. 18. Модель на основе декомпозиции временного ряда
Fig. 18. Model based on time series decomposition

Табл. 7. Оценки качества модели декомпозиции
Table 7. Assessments of the decomposition model

Название	Обозначение	Формула	Значение
Коэффициент детерминации	R^2	$1-\frac{\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$	0,95
Средняя квадратичная ошибка	*MSE*	$\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2$	58 191 261,93
Средняя абсолютная ошибка	*MAE*	$\frac{1}{N}\sum_{i=1}^{n}\left\|a(x_i)-y_i\right\|$	6 533,67
Корень из средней квадратичной ошибки	*RMSE*	$\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(a(x_i)-y_i\right)^2}$	7 628,32
Средняя абсолютная процентная ошибка	*MAPE*	$\frac{1}{N}\sum_{i=1}^{n}\frac{\left\|a(x_i)-y_i\right\|}{\left\|y_i\right\|}\times 100\%$	10,08
Взвешенная абсолютная процентная ошибка	*WAPE*	$\frac{\sum_{i=1}^{n}\left\|Y_i-e_i\right\|}{\sum_{i=1}^{n}\left\|Y_i\right\|}\times 100\%$	9,71

Источник: составлено авторами.

Таким образом, модель с декомпозицией данных продемонстрировала сопоставимое качество с моделью случайного леса. Обе модели показали высокую эффективность, но при сравнении суммы недельных прогнозов модель с декомпозицией оказалась более точной. В ка-

честве вывода отметим, что выбор между моделями зависит от постановки конкретной задачи. Так, в случае, когда требуется короткий прогноз по дням, лучше использовать модель случайного леса, если же необходим прогноз на неделю в целом, более предпочтительной является модель с декомпозицией данных. Иными словами, важно подходить к выбору модели, отталкиваясь от цели прогнозирования и требуемой точности прогнозов.

СПИСОК ИСТОЧНИКОВ

1. *Бондарева К. И.* Понятие и сущность продажи товаров в современных условиях // Экономика и социум. 2016. № 6-3 (25). С. 9–12. EDN WMTGLJ

2. *Зверев О. А.* Система продаж банковских продуктов как неотъемлемый элемент рыночного механизма в банковской сфере // Финансы и кредит. 2004. № 14 (152). С. 3–9. EDN HVQOPL

3. *Чернов М. В.* Понятие и сущность процесса продаж // Экономика и управление: анализ тенденций и перспектив развития. 2016. № 26. С. 76–79. EDN VWSGTD

4. *Плотникова А. В., Хашова В. В., Вишнякова А. Б.* Прогнозирование как элемент принятия управленческих решений в деятельности ПАО «Сбербанк России» // Вестник молодых ученых Самарского государственного экономического университета. 2018. № 2 (38). С. 123–127. EDN VMOAHK

5. *Руденко И. В.* Управление продажами: истоки, сущность, подходы // Вестник Омского университета. Серия: Экономика. 2012. № 4. С. 21–25. EDN QJCIOZ

6. *Мифодовская Ю. С.* Анализ и прогнозирование продаж и закупок на основе математических моделей для торговых компаний // Инновации. Наука. Образование. 2021. № 34. С. 2710–2713. EDN EEMBSQ

7. *Хорзова Я. А.* Применение различных методов прогнозирования объема продаж // Электронный научный журнал. 2016. № 4 (7). С. 596–603. DOI 10.18534/enj.2016.04.596. EDN WAQCOF

8. *Афанасьев Г. И., Афанасьев А. Г., Бурмистрова М. В., Тэт В. Я. С.* Исследование методов машинного обучения для прогнозирования эффективных бизнес-решений в системах электронной коммерции // E-Scio. 2022. № 11 (74). С. 1–14. EDN KCTBIG

9. *Валиахметова Ю. И., Идрисова Э. И.* Применение методов машинного обучения в области прогнозирования объема продаж с учетом динамически изменяющихся признаков // StudNet. 2020. Т. 3, № 10. С. 98. EDN GRCMQK

REFERENCES

1. Bondareva K. I. Ponyatie i sushchnost' prodazhi tovarov v sovremennykh usloviyakh. *Ekonomika i sotsium*, 2016, no. 6-3 (25), pp. 9–12. (In Russ.). EDN WMTGLJ

2. Zverev O. A. Sistema prodazh bankovskikh produktov kak neot"emlemyi element rynochnogo mekhanizma v bankovskoi sfere. *Finance and Credit*, 2004, no. 14 (152), pp. 3–9. (In Russ.). EDN HVQOPL

3. Chernov M. V. Ponyatie i sushchnost' protsessa prodazh. *Ekonomika i upravlenie: analiz tendentsii i perspektiv razvitiya*, 2016, no. 26, pp. 76–79. (In Russ.). EDN VWSGTD

4. Plotnikova A. V., Khashova V. V., Vishnyakova A. B. Prognozirovanie kak element prinyatiya upravlencheskikh reshenii v deyatel'nosti PAO «Sberbank Rossii». *Vestnik molodykh uchenykh Samarskogo gosudarstvennogo ekonomicheskogo universiteta*, 2018, no. 2 (38), pp. 123–127. (In Russ.). EDN VMOAHK

5. Rudenko I. V. Sales management: Origins, essence, approaches. *Herald of Omsk University. Series: Economics*, 2012, no. 4, pp. 21–25. (In Russ.). EDN QJCIOZ

6. Mifodovskaya Yu. S. Analiz i prognozirovanie prodazh i zakupok na osnove matematicheskikh modelei dlya torgovykh kompanii. *Innovatsii. Nauka. Obrazovanie*, 2021, no. 34, pp. 2710–2713. (In Russ.). EDN EEMBSQ

7. Khorzova Ya. A. Primenenie razlichnykh metodov prognozirovaniya ob"ema prodazh. *Elektronnyi nauchnyi zhurnal*, 2016, no. 4 (7), pp. 596–603. (In Russ.). DOI 10.18534/enj.2016.04.596. EDN WAQCOF

8. Afanas'ev G. I., Afanas'ev A. G., Burmistrova M. V., Tet V. Ya. S. Issledovanie metodov mashinnogo obucheniya dlya prognozirovaniya effektivnykh biznes-reshenii v sistemakh elektronnoi kommertsii. *E-Scio*, 2022, no. 11 (74), pp. 1–14. (In Russ.). EDN KCTBIG

9. Valiakhmetova Yu. I., Idrisova E. I. Primenenie metodov mashinnogo obucheniya v oblasti prognozirovaniya ob"ema prodazh s uchetom dinamicheski izmenyayushchikhsya priznakov. *StudNet*, 2020, vol. 3, no. 10, pp. 98. (In Russ.). EDN GRCMQK

10. *Антонов Г. В., Иванов С. И.* Линейная регрессия как один из методов статистического исследования // Известия Великолукской государственной сельскохозяйственной академии. 2021. № 2 (35). С. 64–75. EDN UNIRWN

11. *Ge H., Fang L.* Prediction Model of Physical Goods Sales based on Time Series Analysis // Frontiers in Business, Economics and Management. 2022. Vol. 5, no. 2. P. 90–97.

12. *Zhang Z.* Sales Prediction Based on ARIMA Time Series and Multifactorial Linear Model // Highlights in Science, Engineering and Technology. 2023. Vol. 38. P. 1–8. DOI 10.54097/hset.v38i.5680

13. *Сердинская Ю. А., Мокшин В. В.* Использование методов машинного обучения для оценки прогнозирования продаж товара // Информатика: проблемы, методы, технологии (IPMT-2022): материалы XXII Междунар. науч.-практ. конф. им. Э. К. Алгазинова. Воронеж: Вэлборн, 2022. С. 1062–1068. EDN NXQUYK

14. *Pavlyshenko B. M.* Machine-Learning Models for Sales Time Series Forecasting // Data. 2019. Vol. 4, no. 1. Article 15. DOI 10.3390/data4010015

15. *Zilrahmi M. A. Yu., Putra A. A., Fitri F.* Comparison Fuzzy Time Series Cheng and Ruey Chyn Tsaur Model for Forecasting Sales at Empat Saudara Store // UNP Journal of Statistics and Data Science. 2023. Vol. 1, no. 3. P. 218–225. DOI 10.24036/ujsds%2Fvol1-iss3%2F56

10. Antonov G. V., Ivanov S. I. Linear regression as a method statistical research. *Izvestiya Velikolukskoi gosudarstvennoi sel'skokhozyaistvennoi akademii*, 2021, no. 2 (35), pp. 64–75. (In Russ.). EDN UNIRWN

11. Ge H., Fang L. Prediction Model of Physical Goods Sales based on Time Series Analysis. *Frontiers in Business, Economics and Management,* 2022, vol. 5, no. 2, pp. 90–97.

12. Zhang Z. Sales prediction based on ARIMA time series and multifactorial linear model. *Highlights in Science, Engineering and Technology,* 2023, vol. 38, pp. 1–8. DOI 10.54097/hset.v38i.5680

13. Serdinskaya Yu. A., Mokshin V. V. Ispol'zovanie metodov mashinnogo obucheniya dlya otsenki prognozirovaniya prodazh tovara. *Informatika: problemy, metody, tekhnologii (IPMT-2022),* Voronezh, 2022, pp. 1062–1068. (In Russ.). EDN NXQUYK

14. Pavlyshenko B. M. Machine-learning models for sales time series forecasting. *Data,* 2019, vol. 4, no. 1, Article 15. DOI 10.3390/data4010015

15. Zilrahmi M. A. Yu., Putra A. A., Fitri F. Comparison fuzzy time series Cheng and Ruey Chyn Tsaur Model for Forecasting Sales at Empat Saudara Store. *UNP Journal of Statistics and Data Science,* 2023, vol. 1, no. 3, pp. 218–225. DOI 10.24036/ujsds%2Fvol1-iss3%2F56

СВЕДЕНИЯ ОБ АВТОРАХ

Анастасия Романовна Ермакова – экономический факультет, Пермский государственный национальный исследовательский университет (Россия, 614068, г. Пермь, ул. Букирева, д. 15); ✉ ermakovanastya2015@yandex.ru

Галина Сергеевна Васёва – кандидат экономических наук, доцент кафедры информационных систем и математических методов в экономике, Пермский государственный национальный исследовательский университет (Россия, 614068, г. Пермь, ул. Букирева, д. 15); ✉ vasyova@econ.psu.ru

INFORMATION ABOUT THE AUTHORS

Anastasia R. Ermakova – Faculty of Economics, Perm State University (15, Bukireva st., Perm, 614068, Russia); ✉ ermakovanastya2015@yandex.ru

Galina Sergeevna Vasyova – Candidate of Economic Sciences, Associate Professor at the Department of Information Systems and Mathematical Methods in Economy, Perm State University (15, Bukireva st., Perm, 614068, Russia); ✉ vasyova@econ.psu.ru