УДК 004.048 **DOI** 10.34757/2413-7383.2023.31.4.005

Н. К. Андриевская, Т. В. Мартыненко, Т. А. Васяева
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий национальный технический университет»
283001, г. Донецк, ул. Артёма, 58

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ, КЛАСТЕРНОГО АНАЛИЗА И НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ ПРИ ПРОГНОЗИРОВАНИИ ЗАКУПОЧНЫХ ЦЕН ЛЕКАРСТВ

N. Andrievskaya, T. Martynenko, T. Vasyaeva
Federal State Budgetary Educational Institution of Higher Education
"Donetsk National Technical University"
283001, Donetsk, st. Artyoma, 58

APPLICATION OF STATISTICAL METHODS, CLUSTER ANALYSIS AND NEURAL NETWORK TECHNOLOGIES IN FORECASTING PROCUREMENT PRICES FOR MEDICINES

В статье рассмотрена проблема планирования закупок лекарственных средств аптечной сети. При определении оптимальной закупочной цены возникает необходимость в прогнозировании цен по историческому массиву данных прайс-листов. Проанализированы традиционные статистические подходы к прогнозированию, а также методы на основе нейронной сети. Реализованы четыре метода: метод скользящего среднего; метод случайного леса; метод K-ближайших соседей; нейронная сеть с архитектурой LSTM. На основании ряда метрик произведена оценка качества прогноза тестируемых методов. Проведенные эксперименты показали высокую точность прогнозирования.
Ключевые слова: нейронная сеть, LSTM, закупки, прогнозирование, метод прогноза, точность прогноза.

The article discusses the problem of planning the procurement of medicines in a pharmacy chain. When determining the optimal purchase price, there is a need to forecast prices using a historical array of price list data. Traditional statistical approaches to forecasting, as well as methods based on a neural network, are analyzed. Four methods are implemented: moving average method; random forest method; K-nearest neighbors' method; neural network with LSTM architecture. Based on a number of metrics, the quality of the forecast of the tested methods was assessed. The experiments performed showed high prediction accuracy.
Key words: neural network, LSTM, procurement, forecasting, forecast method, forecast accuracy.

Общая постановка проблемы

Одной из основных функций, требующих автоматизации в отделе продаж аптечной сети, является планирование закупок. Для определения «оптимальной» цены закупки требуется выполнить прогнозирование данных по историческому, т.е. накопленному массиву значений прайсов поставщиков [1].

Задачу краткосрочного и среднесрочного прогнозирования данных по временному ряду возможно решать несколькими способами.

Первая группа прогнозных моделей – модели на основе теории игр модели равновесия по Нэшу, модель Курно, модель Бертрана и др. [2].

Ко второй группе моделей относят так называемые имитационные или фундаментальные модели. С учётом того, что эти модели требуют большого объёма исходных данных, их применение для краткосрочного прогнозирования ограничено.

Третья группа прогнозных моделей основывается на анализе временных рядов с помощью совокупности математико-статистических методов, которые созданы для выявления структуры временных рядов, изучения исторической динамики исследуемых показателей и экстраполяции их на перспективу. В данной группе прогнозных моделей выделяют две подгруппы: традиционные статистические модели и модели искусственного интеллекта.

Использование традиционных статистических моделей позволяет экстраполировать значение временного ряда [3]. Между тем, действительность будет регулярно вносить свои корректировки в аналитическую деятельность. В случае, если система прогнозирования выдаёт абсолютно точную информацию, то трудностей всё равно не избежать. Информационная система использует архивные данные, на основе которых создаётся предположение о том, что произойдет в будущем. Но появление множества значительных перемен на рынке, к сожалению, невозможно предугадать. Следует постоянно помнить, что любой метод прогнозирования несовершенен, так как он не может описать всё многообразие природных и экономических процессов и построить информационную систему, идеально предсказывающую цену, практически невозможно. Некоторые факторы проявляются случайным образом, и установить взаимосвязь между этими факторами, ценой, и коэффициентом влияния этого фактора довольно сложно [4].

На сегодняшний день использование нейронных сетей в различных областях приложения является наиболее популярным решением. К основным преимуществам нейронных сетей относятся высокая гибкость и наиболее точные предсказания значений временных рядов [5]. В отличие от традиционных статистических моделей, которые по существу являются линейными, модели искусственного интеллекта позволяют учесть сложную нелинейную взаимосвязь между зависимой и объясняющими переменными и ассоциировать исследуемый показатель с набором различных факторов, а не только исторических значений.

Так в работе [6] показана возможность использования моделей на основе нейронных сетей для выбора поставщиков услуг, а в работе [7] рассмотрена возможность применения методов с использованием искусственных нейронных сетей для прогнозирования потребности в критических запасных частях.

В статье [8] исследуется эффективность методов машинного обучения для прогнозирования поведения водопроводной сети и построены предсказательные модели значений давления воды в контрольных точках на сети.

Прогнозу спроса на товар с помощью нейронных сетей в условиях меняющейся размерности входных данных посвящены исследования, приведенные в работе [9].

Применение искусственных нейронных сетей для прогнозирования закупок описано в работах [10-13]. В работе [14] автор анализирует сети прямого распространения и делает вывод, что они не подходят для решения поставленной задачи прогнозирования закупок. Там же показано, что сети рекуррентные нейронные сети (RNN) способны обнаруживать только краткосрочные зависимости. Нейронные сети с долгой краткосрочной памятью (LSTM) представляют собой особый тип RNN, которые имеют более длинную «память», чем их предшественники, и способны изучать долгосрочные зависимости. Таким образом, по мнению автора, модель с архитектурой на основе LSTM-ячеек становится лучшей в определении правильной информации, что приводит к «хорошим» прогнозам.

Применение рекуррентных сетей разного вида и LSTM сетей для прогноза цен лекарств в условиях аптечной сети «36,6» описано авторами в работах [15], [16].

Несмотря на достаточное количество исследований, можно сделать вывод, что не существует готового решения для задачи прогнозирования закупочной цены по накопленному массиву прайс-листов, следовательно задачу можно считать актуальной. Таким образом, в процессе исследований необходимо построить прогнозную модель для дальнейшей ее реализации информационной системой (ИС) отдела продаж.

Описание исследуемых методов

А. Метод «скользящего среднего»

Скользящее среднее используется для обнаружения главных тенденций и циклов во временных рядах, а также для сглаживания кратковременных колебаний. Этот метод может быть полезен при краткосрочном прогнозировании, но при этом требует осторожности и внимательности [17]. Преимущества и недостатки метода сведены в табл. 1.

Таблица 1 –Описание метода «скользящего среднего»

Преимущества	Недостатки
1) простота, лёгкость интерпретации; 2) возможность использования для краткосрочного прогнозирования; 3) возможность понятной геометрической интерпретации; 4) подходит для сглаживания ряда; 5) не конфликтность с другими методами	1) укорачивание сглаженного ряда по сравнению с фактическим, что ведет к потере информации; 2) субъективность выбора порядка (периода усреднения); 3) невозможность выражения трендовой тенденции в аналитической форме; 4) «запаздывание» метода (средняя слабо реагирует на резкие развороты); 5) применим только для рядов, которые имеют линейный тренд; 6) использование большого объёма информации

Вывод: метод «скользящего среднего» прост в использовании, но имеет ряд недостатков, что повышает погрешность прогнозов. Поэтому будем применять его в качестве метода для сглаживания временных рядов, а не для прогноза.

B. Метод «экспоненциального сглаживания»

Основывается на методе наименьших квадратов, но с учётом более поздних наблюдений, что позволяет учитывать изменения во времени. Метод сглаживания даёт наиболее точный прогноз на среднесрочную перспективу [18]. При прогнозировании этим методом одним из главных вопросов является выбор оптимального параметра сглаживания *a*. При разных значениях *a* результаты прогноза могут отличаться. Если *a* близка к 1, то в прогнозе учитывается влияние только последних наблюдений, а если а близка к нулю – веса, используемые для взвешивания значений временного ряда, уменьшаются медленно и при прогнозировании учитываются все или почти все наблюдения. Преимущества и недостатки метода сведены в табл. 2.

Таблица 2 – Описание метода «экспоненциального сглаживания»

Преимущества:	Недостатки:
– простота реализации; – используется для среднесрочного прогнозирования; – требует малый объём информации	– сложность подбора параметра сглаживания; – будущий прогноз зависит от точности предыдущего прогноза

Вывод: метод экспоненциального сглаживания полезен при среднесрочных прогнозах, прост в реализации, но имеет значительный недостаток – необходимость подбора параметра сглаживания, влияющего на точность расчёта. Более уместен при использовании в процессе сглаживания временного ряда, чем при прогнозировании.

C. Метод «случайного леса» (Random Forest Regressor)

Благодаря своей гибкости и универсальности, метод «случайного леса» может быть использован для решения различных задач машинного обучения. Однако, при использовании метода «случайного леса» прогнозируемые значения никогда не выходят за пределы значений обучающего набора для целевой переменной. Когда перед алгоритмом ставится задача прогнозирования значений, которые ранее не наблюдались, он всегда будет предсказывать среднее значение ранее просмотренных значений. Очевидно, что среднее значение выборки не может выходить за пределы самых высоких и самых низких значений в выборке. Метод «случайного леса» не сможет обнаружить тенденции, которые позволили бы ему экстраполировать значения, выходящие за пределы обучающей выборки (табл. 3). Столкнувшись с таким сценарием, алгоритм предполагает, что прогноз будет близок к максимальному значению в обучающем наборе [19].

Таблица 3 – Описание метода «случайного леса»

Преимущества:	Недостатки:
– умение эффективно анализировать данные с большим количеством признаков и категорий; – одинаково хорошо анализируются как непрерывные, так и дискретные характеристики; – нечувствительность к масштабированию признаков; – есть методы оценки значимости отдельных признаков; – внутренняя оценка способности модели к обобщению (тест по неотобранным образцам)	– значительный размер получающихся моделей; – прогнозируемые значения никогда не выходят за пределы значений обучающего набора

Вывод: метод лёгок в исполнении, применим на различных наборах данных, но присутствует весомый недостаток – прогнозируемые значения никогда не выходят за пределы значений обучающего набора.

D. Метод «k-ближайших соседей»

Метод «k-ближайших соседей» (*kNN – k Nearest Neighbours*) – это метод решения задач регрессии и классификации, основанный на поиске объектов с известными значениями целевых переменных, которые находятся в непосредственной близости от рассматриваемого объекта. В случае использования метода для решения задачи прогноза рассматриваемому объекту присваивается среднее значение k объектов, чьи значения уже известны и находятся ближе всего к нему. Число k – это количество соседних объектов в пространстве признаков, которые сравниваются с классифицируемым объектом [20]. Преимущества и недостатки метода сведены в табл. 4.

Таблица 4 – Описание метода «k-ближайших соседей»

Преимущества:	Недостатки:
– нужно знать только функцию близости между объектами, сами признаки не нужны; – простая логика работы, легко объяснить и реализовать; – интерпретируемость; – не требует обучения	– высокая сложность одного прогноза; – требования по памяти тоже высоки, т.к. нужно хранить всю выборку; – точность ухудшается с ростом размерности пространства

Вывод: метод прост в реализации, небольшой объём настраиваемых параметров, но имеет недостаток – уменьшение точности прогноза с увеличением выборки.

E. Методы, основанные на нейронных сетях

Для прогнозирования цен и объёмов продаж подходит несколько типов нейронных сетей. Рекуррентные нейронные сети (*Recurrent Neural Networks, RNN*) – это тип многослойного перцептрона, но имеет одну особенность: нейроны могут получать информации не только из предыдущего слоя, но также из предыдущих проходов. Это означает, что порядок подачи данных и обучения сети становится критически важным. Большой трудностью сетей RNN является проблема исчезающего градиента, которая заключается в быстрой потере информации с течением времени [21].

Сети с долгой краткосрочной памятью (*Long Short Term Memory, LSTM*) используют для решения проблемы потери информации путем применения фильтров и чётко определенной ячейки памяти. Каждый нейрон имеет ячейку памяти и три типа фильтров: входной, выходной и фильтр забывания. Цель этих фильтров – обеспечить защиту информации [22]. Фильтр входного типа определяет, какое количество информации из предыдущего уровня будет сохранено в ячейке памяти. Фильтр выходного типа определяет, сколько данных получат последующие слои и выполняет собственно функцию забывания. Обучение нейронной сети зависит от значений весовых коэффициентов. Преимущества и недостатки метода сведены в табл. 5.

Таблица 5 – Описание методов, основанных на нейронных сетях

Преимущества:	Недостатки:
– возможность моделирования не линейных процессов; – адаптивность; – масштабируемость; – разнородность решаемых задач	– сложность программной реализации; отсутствие промежуточных вычислений; – высокие требования к непротиворечивости обучающей выборки

Вывод: метод является самым перспективным и гибким из вышеописанных, может решать различные задачи, но он является относительно трудным в реализации.

Оценка качества

Приведем распространенные меры оценки качества для задач прогноза [23].

Средняя квадратичная ошибка (англ. Mean Squared Error, MSE)

В условиях, когда необходимо выявить значительные ошибки и определить модели, которые дают меньше таких ошибок прогнозирования, может быть применена ошибка MSE:

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad (1)$$

где n – количество наблюдений, по которым строится модель и количество прогнозов; yi – фактические значение зависимой переменной для i-го наблюдения; $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

В этом случае ошибки возводятся в квадрат, что позволяет сделать грубые ошибки более заметными. Мера имеет тенденцию уменьшать качество модели и является чувствительной к аномалиям, а также сложна в интерпретации из-за квадратичной зависимости.

Корень из среднеквадратичной ошибки (англ. Root Mean Squared Error, RMSE)

Мера вычисляется просто как квадратный корень из MSE по формуле 2:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \quad (2)$$

где n – количество наблюдений, по которым строится модель и количество прогнозов; yi – фактические значение зависимой переменной для i-го наблюдения; $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

Эта мера может приводить к снижению качества модели и является чувствительной к аномальным значениям.

Средняя абсолютная ошибка (англ. Mean Absolute Error, MAE)

Рассчитывается как средняя абсолютная разность между наблюдаемыми и предсказанными значениями. При этом, все ошибки взвешиваются одинаково, что позволяет получить более точные результаты. Формула для расчёта MAE:

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad (3)$$

где n – количество наблюдений, по которым строится модель и количество прогнозов; yi – фактические значение зависимой переменной для i-го наблюдения; $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

Мера ошибки проста в понимании и имеет способность уменьшать качество модели, остается чувствительной к выбросам.

Средняя абсолютная процентная ошибка (англ. Mean Absolute Percentage Error, MAPE)

Формула для расчёта MAPE:

$$MAPE = \frac{100}{n} \times \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i|}, \quad (4)$$

где n – количество наблюдений, по которым строится модель и количество прогнозов; yi – фактические значение зависимой переменной для i-го наблюдения; $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

Основная проблема меры – ее нестабильность. Эта мера не имеет размерности и является самой простой для понимания и интерпретации результатов, поскольку измеряется в долях (процентах).

R-квадрат

Коэффициент детерминации (*Coefficient of determination*) или R-квадрат является мерой, которая показывает, насколько хорошо регрессионная модель объясняет дисперсию зависимой переменной. Значение коэффициента детерминации можно использовать для оценки качества модели и её способности прогнозировать результаты. Наиболее общей формулой для вычисления коэффициента детерминации является следующая:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}, \quad (5)$$

где n – количество наблюдений, по которым строится модель и количество прогнозов; yi – фактические значение зависимой переменной для i-го наблюдения; $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

Коэффициент детерминации колеблется в пределах от −∞ до 1. Значения близкие к 1 указывают на полную степень совпадения модели с данными. Если же коэффициент стремится к 0, это означает, что связь между переменными регрессионной модели отсутствует, и вместо нее для оценки значения выходной переменной можно использовать простое среднее ее наблюдаемых значений. В некоторых случаях коэффициент может принимать отрицательные значения, что происходит, когда ошибка модели простого среднего становится меньше ошибки регрессионной модели. Таким образом, добавление в модель с константой некоторой переменной только ухудшает таковую [24]. Применение коэффициента детерминации для оценки качества модели и выбора наилучшей без дополнительных исследований, направленных на доказательство того, что модель подобрана правильно, необоснованно [25].

Разработка и тестирование моделей прогнозирования

Тестовый прогноз производился на 10 дней (*pred_days*) на подготовленном наборе данных. Оценивать качество каждого из алгоритмов будем с помощью метрик, описанных выше.

А. Метод «скользящего среднего»

Для данного метода этап со сглаживанием тренировочного набора данных опускается. Метод тестировался на всём наборе данных (без разделения на train/valid), кроме последних предсказываемых 10 дней.

Основные параметры метода: скользящее окно (time_step) – количество последних дней, используемых для предсказания и в дальнейшем окно, на которое мы сдвигаем скользящую среднюю. В ходе экспериментов были протестированы следующие значения time_step: 5, 10 и 30. Результаты тестирования метода «скользящей средней» представлены в табл. 6.

Таблица 6 – Результаты тестирования метода «скользящего среднего»

time step	**RMSE**		**MSE**		**MAE**		**R2**		**MAPE**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
5	3,96	2,84	15,69	8,066	2,33	2,726	0,997	-11,4	0,015	0,02
10	6,54	2,74	42,78	7,558	3,862	2,591	0,993	-10,58	0,025	0,019
30	11,1	1,66	123,4	2,776	6,752	1,51	0,979	-3,254	0,043	0,011

По приведенным результатам можно сделать выводы:

- увеличение скользящего окна влияет положительно на точность прогноза;
- увеличение скользящего окна уменьшает точность прогноза на тренировочных данных.

Наиболее оптимальным вариантом для тестовых данных стал эксперимент с размером скользящего окна, равным 30.

В. Метод «случайного леса»

Основные параметры метода *Random Forest Regressor*:

- скользящее окно (*time_step*) – количество последних дней, используемых для предсказания и в дальнейшем окно, на которое мы сдвигаем данные.
- n_estimators – число деревьев (чем их больше, тем более высокое качество, однако время настройки и эксплуатации RF также повышается в пропорциональной зависимости).
- max_features – число признаков для выбора расщепления (с увеличением максимального количества функций увеличивается время построения дерева, а деревья получаются «более однообразными». По умолчанию он равен n/3 в задачах регрессии).

Метод тестировался на разделенном наборе данных: train = 90%, valid = 10% и test = 10 дней, а также на наборе train = 90%, valid = 10% и test = 30 дней. В ходе экспериментов на каждом наборе были получены результаты для следующих параметров: time_step=10; n_estimators = 100 и 450; max_features = 30 и 150. Также был проведен эксперимент с обработкой несглаженного ряда. Результаты тестирования метода «случайного» леса выборочно представлены в табл. 7.

Таблица 7 – Результаты тестирования метода «случайного леса»

Параметры	Выборка	RMSE	MSE	MAE	R2	MAPE
time_step = 10 n_estimators = 450 max_features =150 не сглаженный	Train	1,044	1,09	0,638	1	0,004
	Valid	6,002	36,02	3,749	0,992	0,023
	Test	0,875	0,766	0,744	-0,17	0,005
time_step = 10 n_estimators = 450 max_features =150 сглаженный	Train	0,283	0,08	0,177	1	0,001
	Valid	6,472	41,88	4,221	0,991	0,027
	Test	0,841	0,708	0,681	-0,08	0,005
time_step = 10 n_estimators = 100 max_features = 30 сглаженный	Train	0,29	0,084	0,18	1	0,001
	Valid	6,484	42,04	4,222	0,991	0,027
	Test	0,742	0,55	0,624	0,157	0,004

По приведенным результатам можно сделать выводы:

- результат со сглаженными данными лучше, чем с несглаженными данными ряда;
- увеличение числа деревьев с 150 до 450 значительно не изменяет точность результатов;
- изменение скользящего окна с 10 до 30 незначительно влияет на точность прогноза.

Самым оптимальным вариантом стал эксперимент со сглаженными данными для n_estimators = 100 и max_features = 30.

С. Метод K-ближайших соседей

Основные параметры метода:

- скользящее окно (time_step) – количество последних дней, используемых для предсказания и в дальнейшем окно, на которое мы сдвигаем данные.
- n_neighbors – количество ближайших соседей, используемых для предсказания.

Метод тестировался на разделенном наборе данных: train = 90%, valid = 10% и test = 10 дней и test = 30 дней.

В ходе экспериментов были протестированы следующие значения time_step: 5, 10, 30; n_ neighbors = 3, 10, 30. Также проведен эксперимент с обработкой несглаженного ряда.

Результаты тестирования метода «k-ближайших соседей» представлены в табл. 8.

Таблица 8 – Результаты тестирования метода «k-ближайших соседей»

Параметры	Выборка	RMSE	MSE	MAE	R2	MAPE
time_step = 10 n_neighbors = 10 не сглаженный	Train	2,969	8,813	1,857	0,998	0,012
	Valid	11,07	122,6	6,378	0,973	0,04
	Test	1,144	1,308	0,805	-1	0,006
time_step = 10 n_neighbors = 30 сглаженный	Train	3,233	10,45	2,048	0,998	0,013
	Valid	19,24	370	11,92	0,892	0,081
	Test	0,627	0,393	0,534	0,397	0,004
time_step = 5 n_neighbors = 3 сглаженный	Train	0,68	0,463	0,407	1	0,003
	Valid	8,404	70,63	4,967	0,985	0,03
	Test	0,874	0,764	0,763	-0,17	0,005

По приведенным результатам можно сделать выводы:

– увеличение числа соседей приводит к увеличению точности прогноза;

– значение скользящего окна и количества соседей влияют друг на друга.

Более оптимальным вариантом стал эксперимент со сглаженными данными и значениями параметров time_step = 10 и n_ neighbors = 30.

D. Нейронная сеть с архитектурой LSTM

Для построения нейросетевой модели из рассмотренных ранее нейронных сетей была выбрана рекуррентная сеть с краткосрочной памятью LSTM [26].

Основные параметры сети:

– скользящее окно (time_step) – количество последних дней, используемых для предсказания и в дальнейшем окно, на которое мы сдвигаем данные.

– epochs – количество эпох обучения.

– batch_size – количество элементов выборки, с которыми идет работа в пределах одной итерации до изменения весов.

Метод тестировался на разделенном наборе данных: train = 90%, valid = 10% и test = 10 дней.

В ходе экспериментов были протестированы следующие значения time_step: 10, 30; n_ epochs = 10,15; batch_size = 10, 20, 30.

Результаты тестирования нейросетевой прогнозной модели представлены в табл. 9.

Таблица 9 – Результаты тестирования нейросетевой модели

Параметры	Выборка	RMSE	MSE	MAE	R2	MAPE
epochs = 10 batch_size = 10 сглаженный	Train	1,316	1,733	0,984	1,000	0,007
	Valid	6,740	45,423	3,959	0,990	0,023
	Test	1,346	1,813	0,970	-1,777	0,007
epochs = 10 batch_size = 20 сглаженный	Train	1,947	3,789	1,346	0,999	0,008
	Valid	7,783	60,568	4,512	0,982	0,028
	Test	1,732	2,999	1,563	-3,594	0,011
epochs = 15, batch_size = 10 сглаженный, слой Dropout(0.2)	Train	2,949	8,699	2,345	0,998	0,015
	Valid	7,438	55,330	4,341	0,988	0,024
	Test	0,985	0,970	0,901	-0,486	0,006
epochs = 10 batch_size = 10 не сглаженный	Train	3,433	11,785	2,471	0,998	0,016
	Valid	6,844	46,841	4,154	0,990	0,025
	Test	6,393	40,865	5,468	-61,605	0,039
epochs = 10 batch_size = 30 сглаженный	Train	2,221	4,933	1,519	0,999	0,011
	Valid	5,427	29,455	3,218	0,996	0,016
	Test	0,682	0,466	0,571	0,287	0,004

По приведенным результатам можно сделать выводы:

– результат со сглаженными данными лучше, чем с несглаженными;
– увеличение числа эпох при выбранной структуре сети немного улучшает точность прогноза;
– увеличение значения batch_size положительно сказывается на точности прогноза, но точность на обучающей выборке наоборот уменьшается;
– добавление дополнительного слоя Dropout, предназначенного для предотвращения обучения, незначительно увеличивает точность.

Наиболее оптимальным вариантом стал эксперимент со сглаженными данными при следующей архитектуре сети: time_step = 10, epochs = 10 и n_ batch_size = 30.

Обучение сети проводилось с помощью метода fit с переданными оптимальными параметрами. Прогноз осуществлялся с помощью функции predict. Фрагмент обучения с результатами ошибки на каждой эпохе обучения представлен на рисунке 1.

```
history = model.fit(X_train,y_train,validation_data=(X_test,y_test),epochs=10,batch_size=30,verbose=1)
train_predict=model.predict(X_train)
test_predict=model.predict(X_test)

Epoch 1/10
108/108 [==============================] - 5s 22ms/step - loss: 0.0031 - val_loss: 7.1421e-04
Epoch 2/10
108/108 [==============================] - 1s 11ms/step - loss: 5.4277e-05 - val_loss: 7.5334e-04
Epoch 3/10
108/108 [==============================] - 1s 12ms/step - loss: 5.3239e-05 - val_loss: 7.0819e-04
Epoch 4/10
108/108 [==============================] - 1s 11ms/step - loss: 5.2624e-05 - val_loss: 6.6854e-04
Epoch 5/10
108/108 [==============================] - 1s 12ms/step - loss: 5.3931e-05 - val_loss: 6.6709e-04
Epoch 6/10
108/108 [==============================] - 1s 12ms/step - loss: 4.9108e-05 - val_loss: 6.1711e-04
Epoch 7/10
```

Рисунок 1 – Фрагмент обучения нейронной сети

Анализ полученных результатов

Методы являются довольно разнообразными, у каждого есть свои преимущества и недостатки. После проведения экспериментов по четырем алгоритмам прогнозирования можно сделать общий вывод: предварительная обработка входных данных влияет положительно на точность прогнозирования, а слишком большой или слишком маленький параметр скользящего окна (time_step) увеличивает погрешность. В ходе экспериментов была сделана попытка подобрать самые оптимальные параметры для каждого из них. Сравнение лучших вариантов каждого метода между собой представлено в таблице 10.

Таблица 10 – Сравнение результатов методов прогнозирования

Алгоритм прогноза	**RMSE**		**MSE**		**MAE**		**R^2**		**MAPE**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Moving Average	11,11	1,666	123,4	2,77	6,752	1,51	0,979	-3,25	0,043	0,011
Random Forest Regressor	0,29	0,742	0,084	0,55	0,18	0,624	1	0,157	0,001	0,004
K-nearest neighbors	3,23	0,627	10,45	0,39	2,048	0,534	0,998	0,397	0,013	0,004
LSTM	2,22	0,682	4,933	0,47	1,519	0,571	0,999	0,287	0,011	0,004

По результатам видно, что построенные в результате экспериментов конфигурации каждой модели дают достаточно низкую погрешность краткосрочного и среднесрочного прогнозирования. Метод «Random Forest Regressor» имеет проблемы, касающийся невозможности прогнозирования значений, которых не было в обучении, поэтому его использовать не будем. Методы «K-nearest neighbors» и нейронная сеть LSTM показали хороший результат предсказания и могут быть использованы при реализации ИС.

Выводы

Статья посвящена исследованию методов прогнозирования в системах управления закупками лекарственных средств аптечной сети. В результате анализа различных подходов к прогнозированию выделены их преимущества и недостатки. Реализованы четыре метода: метод «скользящего среднего»; метод «случайного леса»; метод «K-ближайших соседей»; нейронная сеть с архитектурой LSTM. На основании ряда мер произведена оценка качества прогноза тестируемых методов. Хотя все методы показали достаточно высокую точность прогнозирования, в качестве выбранной модели следует использовать модель на базе нейронной сети архитектуры LSTM, поскольку ее преимущество подтверждается не только выводами других исследователей, но также проведенными экспериментам.

Список литературы

1. Светличная В.А. Использование методов теории принятия решений для выбора оптимальной стратегии при закупке лекарственных средств / В.А. Светличная, Е.А. Шумаева, О.В. Ченгарь, А.В. Андриевская. *Экономика строительства и городского хозяйства*. 2020. Т. 16. № 1. С. 41-48.
2. Золотова И.Ю. Краткосрочное прогнозирование цен на российском оптовом рынке электроэнергии на основе нейронных сетей. *Проблемы прогнозирования*. 2017.
3. Андриевская, А.В. Экстраполяционные методы прогнозирования закупочных цен лекарств в условиях аптечной сети / А.В. Андриевская, В.О. Вовченко, Н.К. Андриевская. *Информатика, управляющие системы, математическое и компьютерное моделирование* (ИУСМКМ-2021). Материалы XII Международной научно-технической конференции в рамках VII Международного Научного форума Донецкой Народной Республики к 100-летию ДонНТУ. 2021, 169-175.
4. Метод скользящей средней в статистике [Электронный ресурс]. URL: https://www.goodstudents.ru/statistika-zadachi/1144-metod-skolzyashej-srednej.html// (дата обращения: 19.05.2023).
5. Feature selection for time-series prediction in case of undetermined estimation. Khmilovyi S., Skobtsov Yu., Vasyaeva T., Andrievskaya N.В сборнике: Biologically Inspired Cognitive Architectures (BICA) for Young Scientists. Proceedings of the First International Early Research Career Enhancement School (FIERCES 2016). Cham, 2016. С. 85-97.
6. Хаирова, С.М., Хаиров, Б.Г., Шимохин, А.В. Методика работы с поставщиками на основе моделирования работы нейронной сети при решении вопросов выбора поставщиков услуг. *Фундаментальные исследования*. № 7, 2020.
7. Хайрутдинов, М.Р. Применение нейронной сети с обратным распространением и нейронной сети с скрытыми нейронами для прогнозирования потребности в критических запасных частях. *Академическая публицистика*. 2021. №3.
8. Беспалова С.В., Романчук С.М., Ермоленко Т.В., Бондаренко В.И. Построение предсказательных моделей параметров давления воды в водораспределительных сетях с помощью методов машинного обучения. *Проблемы искусственного интеллекта*. 2019. №2 (13).
9. Сазонтьев, В. В. Прогнозирование цен на услуги и товары с использованием нейронных сетей/ Под общей редакцией: Тихонов А. Н., Азаров В. Н., Аристова У. В., Карасев М. В., Кулагин В. П., Леохин Ю. Л., Львов Б. Г., Титкова Н. С. *Научно-техническая конференция студентов, аспирантов и молодых специалистов НИУ ВШЭ.* Материалы конференции/ М. : МИЭМ НИУ ВШЭ, 2014. С. 84-85.

10. Зюсько, К.Д. Прогноз спроса на товар с помощью нейронных сетей в условиях меняющейся размерности входных данных. *Экономика и качество систем связи*. 2020. №1.
11. Рубан, О. И. Использование технологии нейросетей в повседневности. *Летняя школа по искусственному интеллекту 2019* / Кафедра системных исследований МФТИ, Институт проблем искусственного интеллекта ФИЦ ИУ РАН, Российская ассоциация искусственного интеллекта. 4-7 июля 2019 г. Россия, кампус МФТИ.
12. Ступак, А. А. Управление запасами с использованием нейронных сетей. *Управление инвестициями и инновациями*. 2017. № 3. С. 95-103.
13. Бутор, Л. В. Применение искусственных нейронных сетей для прогнозирования закупок = Application of artificial neural networks for procurement forecasting. *Инженерная экономика* [Электронный ресурс] : сборник материалов международной научно-технической конференции профессорско-преподавательского состава в рамках 20-й Международной научно-технической конференции «Наука – образованию, производству, экономике», 26-28 апреля 2022 / Белорусский национальный технический университет, Машиностроительный факультет ; редкол.: А. В. Плясунков, Т. А. Сахнович ; сост. А. В. Плясунков. Минск : БНТУ, 2022. С. 12-15.
14. Использование нейронных сетей в задаче прогнозирования закупок товаров / Д. А. Балавнев, М. Л. Киндулов, Б. Р. Горелов, Т. О. Шергин. *Молодой ученый*. 2020 № 27 (317). С. 30-32.
15. Stock Prices Dynamics Forecasting with Recurrent Neural Networks / T. Vasyaeva, T. Martynenko, S. Khmilovyi, N. Andrievskaya. *Открытые семантические технологии проектирования интеллектуальных систем*. 2020. No 4. P. 277-282.
16. Stock prices forecasting with LSTM networks. Vasyaeva T., Martynenko T., Khmilovyi S., Andrievskaya N. *Communications in Computer and Information Science*. 2019. Т. 1093. С. 59-69.
17. Возможности и недостатки использования скользящей средней при выработке прогнозных решений. *Приоритетные научные направления: от теории к практике*. 2015. №19.
18. Труфанова, Т.В. Нещеменко, К.Д. Способы прогнозирования курса валют на основе моделей экспоненциального сглаживания и Хольта. *Вестник Амурского государственного университета. Серия: Естественные и экономические науки*. 2019. №87.
19. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest) [Электронный ресурс]. URL: https://clck.ru/335YFV// (дата обращения: 19.05.2023).
20. Метод ближайших соседей (kNN) [Электронный ресурс]. URL: https://clck.ru/34YodU (дата обращения: 19.05.2023).
21. Типы нейронных сетей. Принцип их работы и сфера применения [Электронный ресурс]. URL: https://otus.ru/nest/post/1263 // (дата обращения: 19.05.2023).
22. Пустынный, Я.Н. Решение проблемы исчезающего градиента с помощью нейронных сетей долгой краткосрочной памяти. *Инновации и инвестиции*. 2020. №2.
23. Турунцева Марина Юрьевна Оценка качества прогнозов: простейшие методы. *Российское предпринимательство*. 2011. №8-1.
24. Коэффициент детерминации (Coefficient of determination) [Электронный ресурс] – URL: https://wiki.loginom.ru/articles/coefficient-of-determination.html // (дата обращения: 19.05.2023).
25. Кинякин, В.Н., Милевская, Ю.С. Некоторые предостережения по проверке качества модели регрессии с помощью коэффициента детерминации. *Вестник Московского университета МВД России*. 2014. №8.
26. Шпаргалка по разновидностям нейронных сетей. Часть первая. Элементарные конфигурации [Электронный ресурс]. URL: https://tproger.ru/translations/neural-network-zoo-1 // (дата обращения: 20.05.2023).

References

1. Svetlichnaya V.A. Ispol'zovanie metodov teorii prinyatiya reshenij dlya vybora optimal'noj strategii pri zakupke lekarstvennyh sredstv / V.A. Svetlichnaya, E.A. SHumaeva, O.V. CHengar', A.V. Andrievskaya // Ekonomika stroitel'stva i gorodskogo hozyajstva. 2020. – T. 16. № 1. – S. 41-48.
2. Zolotova I.YU. Kratkosrochnoe prognozirovanie cen na rossijskom optovom rynke elektroenergii na osnove nejronnyh setej / I.YU. Zolotova, V.V. // Problemy prognozirovaniya. 2017.
3. Andrievskaya, A.V. Ekstrapolyacionnye metody prognozirovaniya zakupochnyh cen lekarstv v usloviyah aptechnoj seti / A.V. Andrievskaya, V.O. Vovchenko, N.K. Andrievskaya // Informatika, upravlyayushchie sistemy, matematicheskoe i komp'yuternoe modelirovanie (IUSMKM-2021). Materialy XII Mezhdunarodnoj

nauchno-tekhnicheskoj konferencii v ramkah VII Mezhdunarodnogo Nauchnogo foruma Doneckoj Narodnoj Respubliki k 100-letiyu DonNTU. 2021, 169-175

4. Metod skol'zyashchej srednej v statistike [Elektronnyj resurs] – URL: https://www.goodstudents.ru/statistika-zadachi/1144-metod-skolzyashej-srednej.html// (data obrashcheniya: 19.05.2023).
5. Feature selection for time-series prediction in case of undetermined estimation. Khmilovyi S., Skobtsov Yu., Vasyaeva T., Andrievskaya N.V sbornike: Biologically Inspired Cognitive Architectures (BICA) for Young Scientists. Proceedings of the First International Early Research Career Enhancement School (FIERCES 2016). Cham, 2016. S. 85-97
6. Hairova S.M. Metodika raboty s postavshchikami na osnove modelirovaniya raboty nejronnoj seti pri reshenii voprosov vybora postavshchikov uslug // Hairova S.M., Hairov B.G., SHimohin A.V. / FUNDAMENTAL'NYE ISSLEDOVANIYA № 7, 2020
7. Hajrutdinov M.R. Primenenie nejronnoj seti s obratnym rasprostraneniem i nejronnoj seti s skrytymi nejronami dlya prognozirovaniya potrebnosti v kriticheskih zapasnyh chastyah //Akademicheskaya publicistika. 2021. №3.
8. Bespalova S.V., Romanchuk S.M., Ermolenko T.V., Bondarenko V.I. Postroenie predskazatel'nyh modelej parametrov davleniya vody v vodoraspredelitel'nyh setyah s pomoshch'yu metodov mashinnogo obucheniya // Problemy iskusstvennogo intellekta. 2019. №2 (13).
9. Sazont'ev V. V. Prognozirovanie cen na uslugi i tovary s ispol'zovaniem nejronnyh setej/ Pod obshchej redakciej: Tihonov A. N., Azarov V. N., Aristova U. V., Karasev M. V., Kulagin V. P., Leohin YU. L., L'vov B. G., Titkova N. S.// Nauchno-tekhnicheskaya konferenciya studentov, aspirantov i molodyh specialistov NIU VSHE. Materialy konferencii/ M. : MIEM NIU VSHE, 2014. S. 84-85.
10. Zyus'ko K.D. Prognoz sprosa na tovar s pomoshch'yu nejronnyh setej v usloviyah menyayushchejsya razmernosti vhodnyh dannyh // Ekonomika i kachestvo sistem svyazi. 2020. №1
11. Ruban O. I. Ispol'zovanie tekhnologii nejrosetej v povsednevnosti [Tekst] / O.I. Ruban // Letnyaya shkola po iskusstvennomu intellektu 2019 / Kafedra sistemnyh issledovanij MFTI, Institut problem iskusstvennogo intellekta FIC IU RAN, Rossijskaya associaciya iskusstvennogo intellekta. – 4-7 iyulya 2019 g. – Rossiya, kampus MFTI.
12. Stupak, A. A. Upravlenie zapasami s ispol'zovaniem nejronnyh setej / A. A. Stupak // Upravlenie investiciyami i innovaciyami. – 2017. – № 3. – S. 95-103. – DOI 10.14529/iimj170312.
13. Butor, L. V. Primenenie iskusstvennyh nejronnyh setej dlya prognozirovaniya zakupok = Application of artificial neural networks for procurement forecasting / L. V. Butor // Inzhenernaya ekonomika [Elektronnyj resurs] : sbornik materialov mezhdunarodnoj nauchno-tekhnicheskoj konferencii professorsko-prepodavatel'skogo sostava v ramkah 20-j Mezhdunarodnoj nauchno-tekhnicheskoj konferencii «Nauka – obrazovaniyu, proizvodstvu, ekonomike», 26-28 aprelya 2022 / Belorusskij nacional'nyj tekhnicheskij universitet, Mashinostroitel'nyj fakul'tet ; redkol.: A. V. Plyasunkov, T. A. Sahnovich ; sost. A. V. Plyasunkov. – Minsk : BNTU, 2022. – S. 12-15.
14. Balavnev, D. A. Ispol'zovanie nejronnyh setej v zadache prognozirovaniya zakupok tovarov / D. A. Balavnev, M. L. Kindulov, B. R. Gorelov, T. O. SHergin. // Molodoj uchenyj. — 2020. — № 27 (317). — S. 30-32.
15. T. Vasyaeva, Stock Prices Dynamics Forecasting with Recurrent Neural Networks [Tekst] / T. Vasyaeva, T. Martynenko, S. Khmilovyi, N. Andrievskaya // Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nyh sistem. – 2020. – No 4. – P. 277-282.
16. Stock prices forecasting with LSTM networks. Vasyaeva T., Martynenko T., Khmilovyi S., Andrievskaya N. Communications in Computer and Information Science. 2019. T. 1093. S. 59-69.
17. Vozmozhnosti i nedostatki ispol'zovaniya skol'zyashchej srednej pri vyrabotke prognoznyh reshenij // Prioritetnye nauchnye napravleniya: ot teorii k praktike. 2015. №19.
18. Trufanova T.V. Sposoby prognozirovaniya kursa valyut na osnove modelej eksponencial'nogo sglazhivaniya i Hol'ta // T.V. Trufanova, K.D. Neshchemenko / Vestnik Amurskogo gosudarstvennogo universiteta. Seriya: Estestvennye i ekonomicheskie nauki. 2019. №87.
19. Mashinnoe obuchenie dlya nachinayushchih: algoritm sluchajnogo lesa (Random Forest) [Elektronnyj resurs] – URL: https://clck.ru/335YFV// (data obrashcheniya: 19.05.2023).
20. Metod blizhajshih sosedej (kNN) [Elektronnyj resurs] – URL: https://clck.ru/34YodU// (data obrashcheniya: 19.05.2023).
21. Tipy nejronnyh setej. Princip ih raboty i sfera primeneniya [Elektronnyj resurs] – URL: https://otus.ru/nest/post/1263 // (data obrashcheniya: 19.05.2023).
22. Pustynnyj YA.N. Reshenie problemy ischezayushchego gradienta s pomoshch'yu nejronnyh setej dolgoj kratkosrochnoj pamyati // Innovacii i investicii. 2020. №2.

23. Turunceva Marina YUr'evna Ocenka kachestva prognozov: prostejshie metody // Rossijskoe predprinimatel'stvo. 2011. №8-1.
24. Koefficient determinacii (Coefficient of determination) [Elektronnyj resurs] – URL: https://wiki.loginom.ru/articles/coefficient-of-determination.html // (data obrashcheniya: 19.05.2023).
25. Kinyakin V.N. Nekotorye predosterezheniya po proverke kachestva modeli regressii s pomoshch'yu koefficienta determinacii // V.N. Kinyakin, YU.S. Milevskaya / Vestnik Moskovskogo universiteta MVD Rossii. 2014. №8.
26. SHpargalka po raznovidnostyam nejronnyh setej. CHast' pervaya. Elementarnye konfiguracii [Elektronnyj resurs] – URL: https://tproger.ru/translations/neural-network-zoo-1 // (data obrashcheniya: 20.05.2023).

RESUME

N. Andrievskaya, T. Martynenko, T. Vasyaeva

Application of statistical methods, cluster analysis and neural network technologies in forecasting procurement prices for medicines

Background: To solve the problem of determining the optimal purchase price, it is necessary to perform short- and medium-term data forecasting based on a historical array of supplier price list values. It is necessary to select a suitable predictive model for its further implementation in the procurement management system.

Materials and methods: A group of mathematical and statistical methods for identifying the structure of time series, studying the historical dynamics of the indicators under study and extrapolating them into the future is promising for research. In this group of predictive models, there are two subgroups: traditional statistical models and artificial intelligence models, among which it is necessary to make a choice.

Results: Analysis of various approaches to forecasting showed their advantages and disadvantages. Four methods are implemented: moving average method; random forest method; K-nearest neighbors' method; neural network with LSTM architecture. Based on a number of metrics, the quality of the forecast of the tested methods was assessed. All methods showed fairly high prediction accuracy

Conclusion: The results showed that a model based on the neural network of the LSTM architecture should be used as a forecasting model for implementation, since the possibility of its use in forecasting systems is confirmed not only by the experiments performed, but also by the opinions of other researchers.

РЕЗЮМЕ

Н.К. Андриевская, Т.В. Мартыненко, Т.А. Васяева

Применение статистических методов, кластерного анализа и нейросетевых технологий при прогнозировании закупочных цен лекарств

Предпосылки: для решения задачи определения оптимальной цены закупки требуется выполнить краткосрочное и среднесрочное прогнозирование данных по историческому массиву значений прайсов поставщиков. Необходимо подобрать подходящую прогнозную модель для дальнейшей ее реализации в ИС управления закупками**.**

Материалы и методы: Перспективной для исследования является группа математико-статистических методов для выявления структуры временных рядов, изучения исторической динамики исследуемых показателей и экстраполяции их на перспективу. В данной группе прогнозных моделей выделяют две подгруппы: традиционные статистические модели и модели искусственного интеллекта, среди которых необходимо делать выбор.

Результаты: Анализ различных подходов к прогнозированию показал их преимущества и недостатки. Реализованы четыре метода: метод скользящего среднего; метод случайного леса; метод K-ближайших соседей; нейронная сеть с архитектурой LSTM. На основании ряда метрик произведена оценка качества прогноза тестируемых методов. Все методы показали достаточно высокую точность прогнозирования

Заключение: Результаты показали, что в качестве модели прогнозирования для реализации следует использовать модель на базе нейронной сети архитектуры LSTM, поскольку возможность ее применения в системах прогнозирования подтверждается не только проведёнными экспериментами, но и мнениями других исследователей.

Андриевская Наталия Климовна – кандидат технических наук, доцент, выполняет обязанности заведующего кафедрой «Автоматизированные системы управления» Донецкого национального технического университета. Область научных интересов: онтологическое проектирование, семантические технологии, интеллектуальные методы управления. Эл. почта: nataandr@yandex.ru, адрес: 283111, г. Донецк, ул. Кобринской, 19, телефон: +79493349151.

Мартыненко Татьяна Владимировна – кандидат технических наук, доцент кафедры автоматизированных систем управления Донецкого национального технического университета. Область научных интересов: анализ видеоинформации, машинное обучение, современные методы оптимизации, интеллектуальные методы управления. Эл. почта: tatyana.v.martynenko@gmail.com, адрес: 283054, г. Донецк, ул. Политбойцов, 3А/89, телефон: +79494136376.

Васяева Татьяна Александровна – кандидат технических наук, доцент, декан факультета информационных систем и технологий Донецкого национального технического университета. Область научных интересов: машинное обучение, нейросетевое и эволюционное моделирование, методы и системы искусственного интеллекта. Эл. почта: vasyaeva@gmail.com, адрес: 286147, г. Макеевка, кв. «Северный», 22/57, телефон: +79493349176.

Статья поступила в редакцию 14.10.2023.