Научная статья
Original article

ПРОГНОЗИРОВАНИЕ СПРОСА НА ТОВАРЫ СРЕДСТВАМИ МАШИННОГО ОБУЧЕНИЯ

FORECASTING DEMAND FOR PRODUCTS USING MACHINE LEARNING

Пилипенко Артем Юрьевич, студент 2 курса магистратуры по направлению подготовки «Информатика и вычислительная техника» (Искусственный интеллект), ФГБОУ ВО «Иркутский национальный исследовательский технический университет» (664074 Россия, г. Иркутск, ул. Лермонтова, д. 83)

Pilipenko Artem Yuryevich, 2nd year Master's student in the field of study «Informatics and Computer Engineering» (Artificial Intelligence), Federal State Budget Educational Institution of Higher Education «Irkutsk National Research Technical University» (83 Lermontova st., Irkutsk, 664074 Russia)

Аннотация. Статья посвящена применению и сравнению методов машинного обучения в решении задачи прогнозирования спроса на товары. Приводится анализ используемого набора данных, который создан для изучения различных моделей и улучшения качеств прогнозирования. В статье используются наиболее актуальные подходы для прогнозирования временных рядов, такие как модели авторегрессии скользящего среднего, алгоритмы градиентного бустинга и модели глубокого обучения. Представлены графики прогнозов моделей и реальных значений, а также структура модели нейронных сетей и настройки алгоритма градиентного бустинга, благодаря которым удалось добиться минимальной ошибки прогноза среди всех задействованных алгоритмов. Практическая значимость проведенного

исследования заключается в том, что результаты и материалы данной работы могут использоваться при прогнозировании спроса на схожем наборе данных.

Annotation. The article is devoted to the application and comparison of machine learning methods in solving the problem of forecasting the demand for goods. An analysis of the data set used, which was created to study various models and improve the quality of forecasting, is given. The article uses the most relevant approaches for time series forecasting, such as moving average autoregressive models, gradient boosting algorithms, and deep learning models. The graphs of model predictions and real values are presented, as well as the structure of the neural network model and the settings of the gradient boosting algorithm, thanks to which it was possible to achieve the minimum prediction error among all involved algorithms. The practical significance of the study is that the results and materials of this work can be used to predict demand for a similar data set.

Ключевые слова: методы машинного обучения, прогнозирование спроса, точность прогнозирования, методы временных рядов, прогноз продаж, визуализация данных, модели авторегрессии, градиентный бустинг, глубокое обучение, нейронные сети.

Keywords: machine learning methods, demand forecasting, forecasting accuracy, time series methods, sales forecasting, data visualization, autoregressive models, gradient boosting, deep learning, neural networks.

Введение

Прогнозирование спроса на товары играет огромную роль и является одним из ключевых процессов в создании технологичных продуктов, улучшении пользовательского опыта и эффективного функционирования организации. При помощи повышения точности прогнозирования можно минимизировать риски, затраты, а также повысить эффективность процессов оптимизации товарных запасов, выстроить план работы компании и товаров, которые наиболее актуальны среди потребителей.

Точность прогноза зависит от используемой модели и самих данных. Когда лежащие в основе механизмы прогнозирования неизвестны, слишком сложны для

понимания или известны не полностью, как розничные продажи, применяют простую статистическую модель. Популярными классическими методам, относящимися к этой категории, являются ARIMA, методы экспоненциального сглаживания, такие как Holt-Winters, и метод Theta, который используется менее широко.

Подходы к компьютерному обучению, в том числе алгоритмы случайного леса, стали частью инструментария прогнозиста. Свою эффективность также показали рекуррентные нейронные сети в случае достаточного количества данных [1].

Невозможно изначально знать, какая модель приведет к оптимальной производительности и выполнит поставленную задачу. Поэтому необходимо применить различные методы прогнозирования спроса, чтобы изучить их особенности и выявить наиболее эффективные среди них для решения текущей задачи.

Набор данных для прогнозирования продаж

Для обучения и тестирования методов машинного обучения использовался набор данных соревнования «Store Item Demand Forecasting Challenge» [2] проходящий в системе организации конкурсов по исследованию данных Kaggle. Данное соревнование позволяет изучить различные методы временных рядов на относительно простом и чистом наборе данных, который содержит информацию о продажах товаров в магазинах за 5 лет. Цель соревнования – спрогнозировать продажи 50 различных товаров в 10 разных магазинах на 3 месяца вперёд.

Набор данных состоит из 3-х таблиц: «train.csv» – тренировочные данные, «test.csv» – тестовые данные и «sample_submission.csv» – образец файла представления результатов. Данные состоят из следующих столбцов: «date» – дата продажи товара, «store» – идентификатор магазина, «item» – идентификатор товара, «sales» – количество товаров, проданных в определённом магазине в определённую дату. Структура таблиц тренировочных и тестовых данных представлена на рисунке 1.

	date	store	item	sales
0	2013-01-01	1	1	13
1	2013-01-02	1	1	11
2	2013-01-03	1	1	14
3	2013-01-04	1	1	13
4	2013-01-05	1	1	10
...	...	...	...	...
912995	2017-12-27	10	50	63
912996	2017-12-28	10	50	59
912997	2017-12-29	10	50	74
912998	2017-12-30	10	50	62
912999	2017-12-31	10	50	82

913000 rows × 4 columns

	id	date	store	item
0	0	2018-01-01	1	1
1	1	2018-01-02	1	1
2	2	2018-01-03	1	1
3	3	2018-01-04	1	1
4	4	2018-01-05	1	1
...	...	...	...	...
44995	44995	2018-03-27	10	50
44996	44996	2018-03-28	10	50
44997	44997	2018-03-29	10	50
44998	44998	2018-03-30	10	50
44999	44999	2018-03-31	10	50

45000 rows × 4 columns

Рисунок 1 – Структура набора данных

Тренировочные данные содержат 913000 строк и 4 столбца. Тестовые данные содержат 45000 строк и 4 столбца. Столбец «id» в тестовых данных необходим для связи с образцом файла представления результатов.

Визуализируем данные используя метод декомпозиции временных рядов [3], который позволяет разложить временные ряды на три отдельные компоненты: тренд, сезонность и шум. На рисунке 2 представлен результат декомпозиции временных рядов.

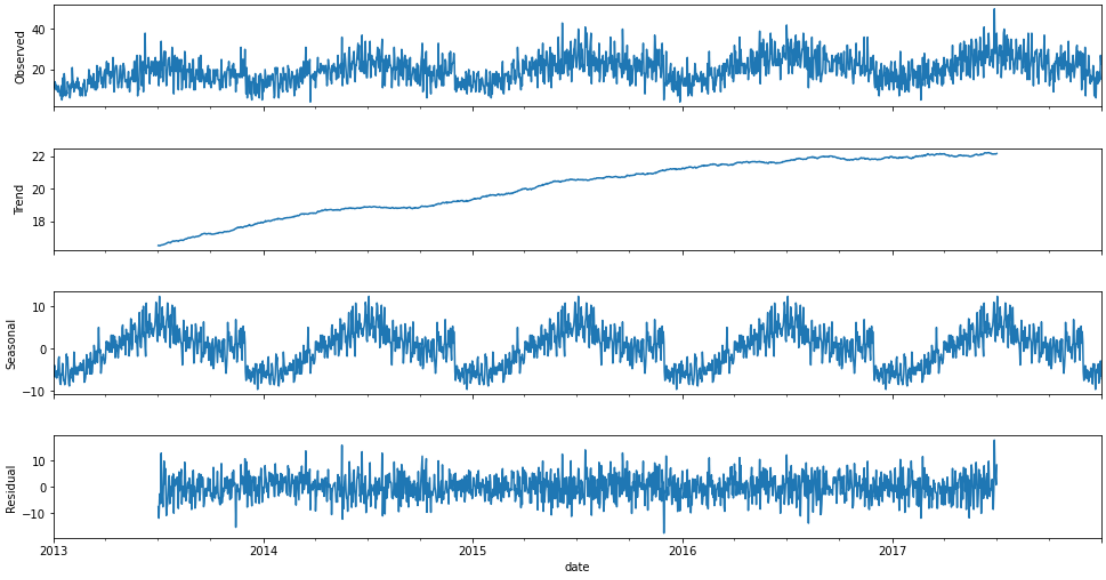


Рисунок 2 – Результат декомпозиции временных рядов

На рисунке 2 видно, что временной ряд нестабилен и имеет сезонный характер, продажи всегда низкие в начале года, высокие в середине года, а в конце года снова низкий уровень продаж. График тренда показывает, что продажи имеют восходящий тренд. Это означает, что данные не являются стационарными, поскольку среднее значение увеличивается со временем.

Модели прогнозирования временных рядов

Ниже представлены модели, которые были выбраны для прогнозирования временных рядов и сравнениях их между собой.

Модели авторегрессии скользящего среднего:

- ARIMA;
- SARIMA;
- ARIMAX;
- SARIMAX.

Модели машинного обучения, использующие алгоритм градиентного бустинга:

- CatBoost;
- LightGBM;
- XGBoost.

Модели глубокого обучения:

- XGBoost.
- Многослойный персептрон (англ. Multilayered perceptron, MLP);
- Свёрточная нейронная сеть (англ. Convolutional neural network, CNN);
- Долгая краткосрочная память (англ. Long short-term memory, LSTM).

Для оценки качества прогноза были взяты следующие метрики:

- Средняя квадратичная ошибка (англ. Mean Squared Error, MSE);
- Средняя абсолютная ошибка (англ. Mean Absolute Error, MAE);
- Коэффициент детерминации;
- Средняя абсолютная процентная ошибка (англ. Mean Absolute Percentage Error, MAPE);

- Корень из средней квадратичной ошибки (англ. Root Mean Squared Error, RMSE).

Обучение проводилось на данных из файла «train.csv», в которых даты продажи с 2013-01-01 по 2017-09-30 будут использоваться для обучения моделей прогнозирования, а с 2017-10-01 по 2017-12-31 для тестирования, так как главная задача соревнования – это спрогнозировать продажи на 3 месяца вперёд, но для оценки моделей необходимо иметь реальное значение продаж.

Вместе с этим рассматривалось время, затраченное моделями на обучение.

Сравнение моделей авторегрессии

Процесс обучения производился в облачном сервисе Google Colaboratory [4], который предоставляет виртуальную машину со следующими характеристиками: 13 ГБ оперативной памяти, 2-х ядерный процессор Intel Xeon 2.20GHz и 66 ГБ жёсткого диска.

В таблице 1 представлены результаты оценки качества прогноза используя перечисленные выше модели авторегрессии скользящего среднего. Так как данные модели могут работать только с одномерными массивами, то изначально обучение производилось на данных первого магазина и первого товара для сравнения этих моделей между собой.

Таблица 1 – Оценки моделей авторегрессии

Модель	MSE	MAE	Коэффициент детерминации	MAPE	RMSE	Время обучения
ARIMA	44.069	5.421	-0.299	36.7 %	6.638	8 сек.
SARIMA	29.624	4.583	0.126	28.9 %	5.442	25 сек.
ARIMAX	24.359	3.994	0.281	25.0 %	4.935	20 сек.
SARIMAX	29.354	4.334	0.134	26.5 %	5.418	36 сек.

Исходя из результатов таблицы 1 можно выделить модель ARIMAX, которая показала лучший результат среди моделей авторегрессии.

Результаты моделей прогнозирования временных рядов

Далее будет использоваться модель ARIMAX на всех данных и оставшиеся модели. В таблице 2 представлены результаты оценки прогнозирования временных рядов, используя все перечисленные выше модели.

Таблица 2 – Результаты оценки моделей прогнозирования временных рядов

Модель	MSE	MAE	Коэффициент детерминации	MAPE	RMSE	Время обучения
ARIMAX	127.859	8.633	0.842	18.9%	11.307	2 ч. 46 мин.
CatBoost	64.330	6.145	0.920	13.5%	8.020	18 мин.
LightGBM	83.867	6.900	0.896	13.9%	9.157	3 сек.
XGBoost	67.013	6.381	0.917	14.5%	8.186	1 ч. 8 мин.
MLP	272.151	14.602	0.665	30.6%	16.497	4 мин.
CNN	68.137	6.357	0.916	13.1%	8.254	15 мин.
LSTM	55.823	5.770	0.931	12.6%	7.471	28 мин.

Все модели обучались без добавления новых признаков и с использованием параметров по умолчанию.

Наибольшее время обучения модели ARIMAX обусловлено тем, что для каждого магазина и товара необходимо создавать и обучать свою модель.

Наилучший результат среди моделей машинного обучения, использующих алгоритм градиентного бустинга, показал CatBoost [5]. XGBoost по точности близок, но время обучения длилось почти в 4 раза больше.

Самый точный результат смогла построить сеть LSTM, которая хорошо приспособлена к прогнозированию временных рядов [6]. На рисунке 3 представленные структуры моделей глубокого обучения, использованных для прогнозирования.

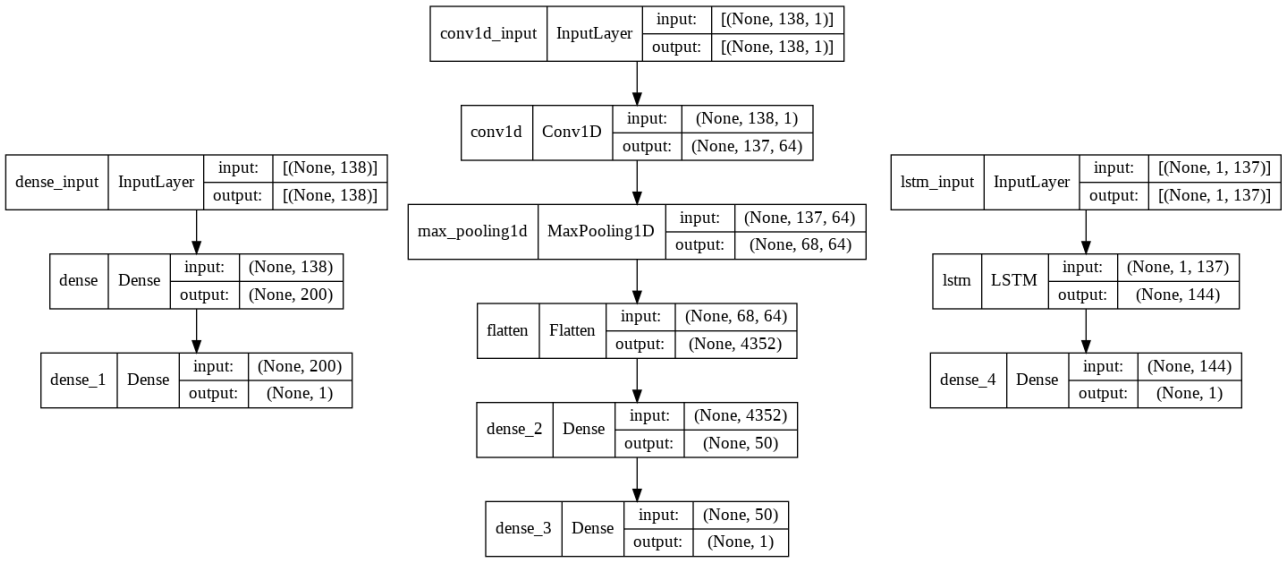


Рисунок 3 – Структуры моделей нейронных сетей MLP, CNN и LSTM

На рисунке 4 представлены графики сравнения прогноза и реальных значений моделей CatBoost и LSTM.

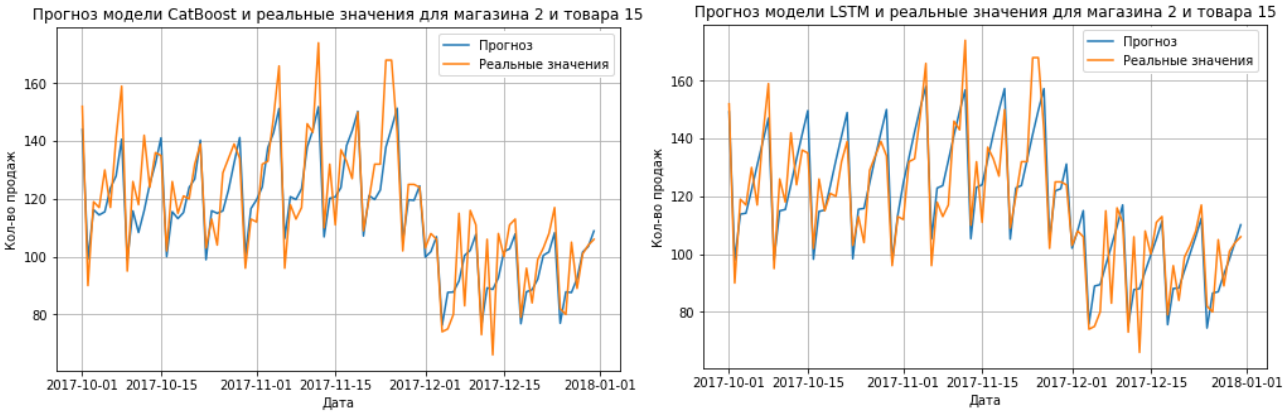


Рисунок 4 – Прогнозы моделей CatBoost и LSTM в сравнении с реальными значениями

Модели смогли предсказать небольшой подъём продаж в ноябре и падение продаж в декабре.

Результаты модифицированных моделей CatBoost и LSTM

На основе проведенных экспериментов, описанных выше, для дальнейшего исследования были выбраны модели – CatBoost и LSTM. Чтобы добиться лучшего результата данные и сами модели необходимо было дополнить и настроить. К данным для обучения модели CatBoost были добавлены дополнительные признаки (см. таблицу 3), которые позволили модели реагировать на средние значения продаж для корректировки прогноза.

Таблица 3 – Дополнительные признаки для улучшения качества прогноза

Название признака	Описание признака
item_quarter_mean	Среднее количество продаж товара за квартал
store_quarter_mean	Среднее количество продаж магазином за квартал
store_item_quarter_mean	Среднее количество продаж товара магазином за квартал
item_month_mean	Среднее количество продаж товара в месяц
store_month_mean	Среднее количество продаж магазином в месяц
store_item_month_mean	Среднее количество продаж товара магазином в месяц
item_weekofyear_ mean	Среднее количество продаж товара в неделю
store_weekofyear_mean	Среднее количество продаж магазином в неделю
store_item_ weekofyear_mean	Среднее количество продаж товара магазином в неделю
itemdayofweek_mean	Среднее количество продаж товара в день недели
storedayofweek_mean	Среднее количество продаж магазином в день недели
storeitemdayofweek_mean	Среднее количество продаж товара магазином в день недели

Обучение модели CatBoost производилось со следующими параметрами: 1000 итераций, глубина дерева 8 и скорость обучения 0.14.

Добавление количества нейронов и слоёв к структуре модели LSTM, представленной на рисунке 3, не улучшило результат. Поэтому было решено изменить подход к обучению и саму структуру модели [7]. Структура новой модели LSTM представлена на рисунке 5.

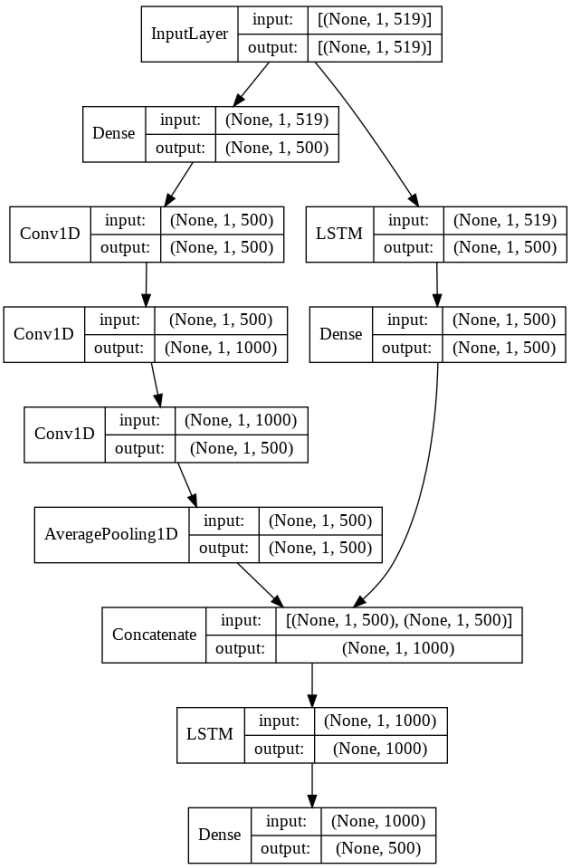


Рисунок 5 – Измененная структура модели LSTM

Обучение производилось на данных, которые содержат: день недели, месяц, и количество проданных товаров в предыдущий день. Прогноз строится по следующему принципу: первая строка содержит количество проданных товаров в предыдущий день (т.е. количество продаж в последний день обучающей выборки) и на основе этого числа модель генерирует количество продаж на все последующие дни.

В таблице 4 представлены результаты оценки прогнозов настроенных моделей CatBoost и LSTM.

Таблица 4 – Результат модифицированных моделей CatBoost и LSTM

Модель	MSE	MAE	Коэффициент детерминации	MAPE	RMSE	Время обучения
CatBoost	54.602	5.708	0.932	12.6%	7.389	11 мин.
LSTM	82.899	6.892	0.898	13.6%	9.104	1 ч. 53 мин.

Добавление новых признаков позволило повысить качество прогноза модели CatBoost, а у модели LSTM ухудшился результат, в сравнении с предыдущим подходом.

На рисунке 6 представлены графики сравнения прогноза и реальных значений модифицированных моделей CatBoost и LSTM.

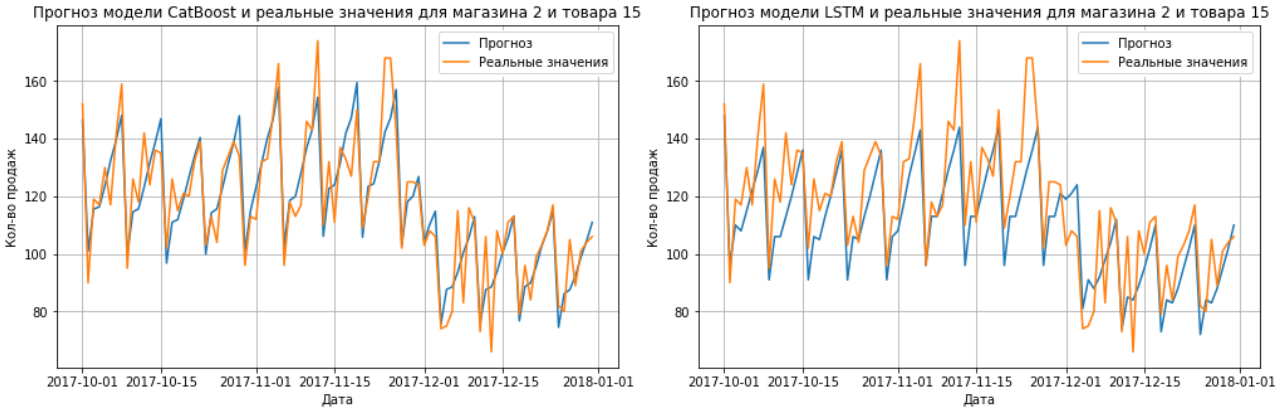


Рисунок 6 – Прогнозы модифицированных моделей CatBoost и LSTM в сравнении с реальными значениями

Заключение

Среди моделей авторегрессии скользящего среднего ARIMAX показала наименьшую ошибку, однако эти модели принимают на вход только данные одного товара и магазина. Это значит, что для решения поставленной задачи необходимо создавать для каждого товара и магазина свою уникальную модель, что замедляет процесс построения прогноза и вынуждает хранить большое количество моделей.

Модели машинного обучения, использующие алгоритм градиентного бустинга, без каких-либо настроек и добавлений новых признаков показывают более точный прогноз в сравнении с моделями авторегрессии. Преимущество модели CatBoost заключается в том, что при использовании категориальных признаков не нужно использовать One-Hot Encoding, а достаточно лишь указать столбцы, содержащие категориальные признаки, что позволяет ускорить процесс обучения.

Модель LSTM показала наименьшую ошибку прогноза при использовании изначальных данных. При использовании данных продаж за предыдущей день и построении прогноза на последующие дни модель обучалась большее количество времени и при этом ошибка прогноза увеличилась.

В результате применения различных моделей для прогнозирования продаж на наборе данных «Store Item Demand Forecasting Challenge» наилучший результат показали две модели – CatBoost с добавлением дополнительных признаков и LSTM с использование изначальных данных. Оценки результатов прогнозов этих моделей получились почти равными. Время обучения для моделей CatBoost и LSTM составило 11 минут и 28 минут соответственно. За это время модели обработали 867000 строк данных. Ошибка по метрике MAPE составила 12,6% от фактических значений. Коэффициент детерминации близок к единице и составляет 0.932 – это говорит о том, что модели хорошо объясняют данные. Результаты проведенных исследований могут быть использованы при прогнозировании спроса на схожем наборе данных.

Литература

1. Прогнозирование: какие методы использует Uber [Электронный ресурс]. – Режим доступа: https://neurohive.io/ru/osnovy-data-science/prognozirovanie-uber/. – (дата обращения: 10.01.2022).
2. Store Item Demand Forecasting Challenge | Kaggle [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/c/demand-forecasting-kernels-only/overview. – (дата обращения: 11.01.2022).
3. How to Decompose Time Series Data into Trend and Seasonality [Электронный ресурс]. – Режим доступа: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/. – (дата обращения: 12.01.2022).
4. Добро пожаловать в Colaboratory! [Электронный ресурс]. – Режим доступа: https://colab.research.google.com/. – (дата обращения: 14.01.2022).
5. CatBoost – open-source gradient boosting library [Электронный ресурс]. – Режим доступа: https://catboost.ai/. – (дата обращения: 16.01.2022).
6. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras [Электронный ресурс]. – Режим доступа: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/. – (дата обращения: 17.01.2022).

7. LSTM Fully Convolutional Networks for Time Series Classification [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1109/ACCESS.2017.2779939. – (дата обращения: 19.01.2022).

References

1. Forecasting: what methods does Uber use [Electronic resource]. – Access mode: https://neurohive.io/ru/osnovy-data-science/prognozirovanie-uber/. – (date of access: 01/10/2022).
2. Store Item Demand Forecasting Challenge | Kaggle [Electronic resource]. – Access mode: https://www.kaggle.com/c/demand-forecasting-kernels-only/overview. – (date of access: 01/11/2022).
3. How to Decompose Time Series Data into Trend and Seasonality [Electronic resource]. – Access Mode: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/. – (date of access: 01/12/2022).
4. Welcome to Collaboratory! [Electronic resource]. – Access mode: https://colab.research.google.com/. – (date of access: 01/14/2022).
5. CatBoost – open-source gradient boosting library [Electronic resource]. – Access mode: https://catboost.ai/. – (date of access: 01/16/2022).
6. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras [Electronic resource]. – Access mode: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/. – (date of access: 01/17/2022).
7. LSTM Fully Convolutional Networks for Time Series Classification [Electronic resource]. – Access Mode: http://dx.doi.org/10.1109/ACCESS.2017.2779939. – (date of access: 01/19/2022).

© Пилипенко А.Ю., 2022 Научно-образовательный журнал для студентов и преподавателей «StudNet» №2/2022.

Для цитирования: Пилипенко А.Ю. ПРОГНОЗИРОВАНИЕ СПРОСА НА ТОВАРЫ СРЕДСТВАМИ МАШИННОГО ОБУЧЕНИЯ// Научно-образовательный журнал для студентов и преподавателей «StudNet» №2/2022.