ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА СОВРЕМЕННЫХ ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

М. Н. Ткачева

Саратовский национальный исследовательский государственный университет им. Н.Г. Чернышевского, Россия
E-mail: tkacheva.MN@yandex.ru

В предложенной работе рассматриваются, как исторические аспекты использования машинного обучения в сфере экономики, так и перспективные направления применения ML. Эпоха, ручного подсчета и человеческого фактора, постепенно уходит на второй план. На первый же план приходят технологии искусственного интеллекта, имеющие куда большие возможности для обработки больших массивов данных. При помощи машинного обучения и его методов, мы можем добиться автоматизации экономических процессов, а также повышения их точности.

USING MACHINE LEARNING TO ANALYSE MODERN ECONOMIC PROCESSES

M. N. Tkacheva

The proposed paper considers both historical aspects of using machine learning in the field of economics and promising directions of ML application. The era of manual counting and human factor is gradually fading into the background. In the first place come the technologies of artificial intelligence, which have much more opportunities for processing large data arrays. With the help of machine learning and its methods, we can achieve automation of economic processes and increase their accuracy.

В современном мире особое значение отводится развитию информационных технологий (ИТ). Экономика, как сложная и динамичная система, постоянно сталкивается с новыми вызовами и возможностями. В эпоху информационных технологий и больших данных, машинное обучение (англ. Machine Learning / ML) (МО) предлагает революционный подход к анализу экономических процессов. МО как инструмент способно обрабатывать огромные массивы данных, выявлять скрытые закономерности, прогнозировать будущие тенденции и оптимизировать экономические решения.

Понятие «машинное обучение» возникло в 1959 г., когда А. Сэмюэль, работающий в IBM, использовал этот термин для описания задач распознавания образов в ранних системах искусственного интеллекта (ИИ). Хотя концепция ИИ начали обсуждаться с 1930-х гг., её систематическое изучение началось после Дартмутского семинара 1956 г., где Дж. Маккарти предложил термин «искусственный интеллект», заменив им «кибернетику». [1] В начале своего развития МО рассматривалось как часть более широкой области ИИ. Однако с течением времени спектр практических применений МО значительно расширился, выйдя за рамки, установленные ИИ. Сегодня МО используется как компонент систем

© Ткачева М. Н., 2024

ИИ и как самостоятельные системы, коих значительно больше [2]. Часто термины «ИИ» и «МО» путают, используя как синонимы. Чтобы избежать путаницы, можно использовать простое правило: если система без вмешательства человека, то это, скорее всего, ИИ. Если же система классифицирует или прогнозирует на основе обучения, то это МО.

Использование МО в экономике имеет давнюю историю. Первой работой, где МО применялось к экономическим задачам, стало исследование «Wang et al», опубликованное в 1984 г. В 1988 г. Х. Уайт использовал нейронные сети для прогнозирования ежедневной доходности акций IBM. С этого момента применение МО в экономике неуклонно растет. Первоначально оно применялось для прогнозирования финансовых временных рядов, где используются большие объёмы данных [1]. В прошлом эффективность МО зависела от наличия огромных наборов данных, что ограничивало его применение в других областях экономики. Кроме того, обучение моделей занимало много времени из-за недостаточной вычислительной мощности. Сегодня появились новые модели МО, которые могут эффективно работать с небольшими наборами данных. Это открывает новые возможности для применения МО в макро и микроэкономике, где данные часто ограничены. Недавние исследования в области прогнозирования экономических колебаний показывают, что МО превосходит традиционные экономические модели.

В настоящее время можно выделить несколько преимуществ МО. Первое – легкость в выявлении трендов и моделей: МО, являющееся частью ИИ, позволяет устройствам учиться самостоятельно, выявляя закономерности и тенденции в данных. ИИ-системы, использующие машинное обучение, находят решения, анализируя эти закономерности. Это делает процессы более эффективными и ускоряет их выполнение, что особенно важно в ситуациях, где время играет ключевую роль. Второе – улучшение навыков в процессе работы: современные устройства легко настраиваются, благодаря специальным ПО. Машинное обучение позволяет избежать постоянных обновлений, так как система сама учится и корректируется, основываясь на текущих тенденциях. Третье – самодостаточность и разнообразие применения: гибкость применения и быстрая обучаемость позволяет применять ML в любой сфере, что делает его более функциональным.

Среди недостатков МО следует отметить: частые ошибки и продолжительное время работы. Алгоритмы МО, созданные людьми, могут содержать ошибки, что приводит к неточностям в результатах. Это особенно опасно в отраслях, где важна точность и обрабатываются большие объёмы данных, например, в производстве. Небольшая ошибка в алгоритме может привести к браку продукции. Поэтому, важно контролировать систему и вмешиваться в ее работу, пока она не будет полностью откалибрована и ошибки устранены. Кроме того, разработка технологий МО требует значительных инвестиций. От создания алгоритмов до обучения специалистов и приобретения специализированного оборудования – каждый этап требует финансовых вложений. Это делает внедрение машинного обучения дорогостоящим процессом.

Не смотря на вышеперечисленные «минусы», МО используют наряду с

традиционными эконометрическими инструментами для улучшения понимания экономических систем. Объединив сильные стороны обеих областей, исследователи могут повысить точность и надежность экономических анализов для лучшего информирования о политических решениях [3]. Стоит также отметить, что привлекательность МО заключается и в том, что ему удается обнаружить обобщающие закономерности. Успех МО в решении экономических задач во многом объясняется его способностью обнаруживать сложную структуру, которая не была задана заранее [4]. МО удается подгонять сложные и очень гибкие функциональные формы к данным без перебора.

Существует множество различных моделей МО. Для иллюстрации использования МО ниже будут приведены несколько моделей. Одна из таких – предиктивная модель для обнаружения мошенничества с кредитными картами. По данным отчета «Fraud the Facts 2019», потери британских компаний, связанные с кражей данных банковских карт, в 2017-2018 гг., достигли 1% общей суммы убытков [5]. В США мошенничество в сфере здравоохранения и страхования приводит к ущербу в 98 млрд и 300 млрд долл. в год [6].

Алгоритм применения упомянутой предиктивной модели следующий: при ее обучении используется набор данных, содержащий транзакции, совершенные по кредитным картам (в сентябре 2013 г. в Европе) [7]. Общее количество транзакций – 284807, из них 492 совершенны в ходе мошенничества. Все транзакции были совершены в течение 2 дней. Данные имеют набор обезличенных характеристик: V1, V2, ... V28, – это главные компоненты, полученные с помощью PCA (один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации), единственные характеристики, которые не были преобразованы с помощью PCA, – это «Время» (секунды, прошедшие между каждой транзакцией и первой транзакцией в наборе данных) и «Сумма» (транзакций). В ходе подготовки к использованию прогностической модели была проведена проверка на несбалансированность, визуализация признаков, выявление взаимосвязи между этими признаками. Данные были разделены на три части: обучающий набор, набор для проверки и тестовый набор. Были опробованы 5 прогностических моделей. Для каждой модели был получен AUC-скалярная величина, показывающая производительность модели:

1) RandomForrestClassifier показатель AUC в 0,85 при прогнозировании цели на тестовом наборе;

2) AdaBoostClassifier показатель AUC в 0,83 при прогнозировании цели на тестовом наборе;

3) CatBoostClassifier показатель AUC в 0,86 после 500 итераций;

4) XGBoost показатель AUC в 0,974 при использовании валидационного набора;

5) LightGBM показатель AUC в 0,946 при прогнозировании цели на тестовом наборе.

Полученный разброс в величинах производительности показывает, что правильная подготовка данных и грамотный подбор модели для дальнейшей обработки статистики, играет ключевую роль [8].

Ротация кадров – одна из важных экономическим проблем, затрагивающая многие отрасли народного хозяйства, влекущая остановку или замедление работы, а также потерю финансов. Чтобы улучшить удержание сотрудников и планировать найм, нужно понять, почему и когда люди уходят. Для этого разработана модель МО «кадровый аналитик для предприятий». В ней используется набор данных от компании IBM, в котором содержится информация о сотрудниках за 2016 г. Ее данные: количество рабочих часов, уровень образования, отношения на работе и т.д., – несут в себе много полезного [9]. После предварительной обработки статистики, для удобства использования и восприятия к данным были применены 6 разных моделей с разным процентом эффективности и корректности. В текущим исследовании каждая модель была оценена по ряду параметров:

1) ROC AUC – мера, которая позволяет суммировать производительность модели одним числом, измеряя площадь под кривой ROC. Использовалось среднее и ожидаемое значение;

2) Accuracy – доля правильно предсказанных классов ко всем предсказаниям. Использовалось среднее и ожидаемое значение (см. табл.). Таким образом, модели были расположены не только по их производительности, но и по величине их правильности ответов. Наилучшим образом себя показала модель «случайный лес», относящаяся к классификационным моделям.

Результаты тестирования алгоритмов МО

	Algorithm	ROC AUC mean	ROC AUC std	Accuracy mean	Accuracy std
0	Logistic Regression	82.03	8.06	74.49	5.53
2	SVM	78.88	8.21	84.48	4.18
1	Random Forest	78.86	7.01	85.30	3.75
5	Gaussian NB	75.06	5.10	68.14	3.14
3	KNN	66.42	9.90	84.21	4.04
4	Decision Tree Classifier	58.02	6.23	76.22	4.23

Примечание. Здесь ROC AUC – производительность МО,
Accuracy – корректность МО, mean – среднее, std – ожидаемое значение,
Algorithm – название алгоритма МО.

На основе исследования была получена модель, прогнозирующая показатель риска ухода сотрудника из компании, и полезные данные о том, какие меры можно предпринять для предотвращения ухода сотрудников.

Сегодня всё больше людей интересуются торговлей на бирже, и некоторые пользователи смогли применить МО в этой сфере. Мы рассматриваем возможность использования моделей МО для анализа временных рядов и прогнозирования их дальнейшей динамики. В качестве моделей для прогнозирования будущих значений временных рядов были использованы: Prophet и Auto-ARIMA.

Prophet – это процедура прогнозирования временных рядов, основанная на аддитивной модели, в которой нелинейные тренды подгоняются под годовую, недельную и дневную сезонность, а также эффекты праздников. Лучше всего она работает с временными рядами, имеющими сильные сезонные эффекты и несколько сезонов исторических данных. Prophet устойчив к недостающим данным и сдвигам в тренде, и обычно хорошо справляется с выбросами. Цель модели ARIMA (Auto Regessive Integrated Moving Average) – предсказать будущие движения ценных бумаг или финансового рынка, изучая разницу между значениями в ряду [10].

Поиск преимуществ и недостатков ИИ и МО далек от завершения. Отметим, что у прогностических моделей, используемых в трейдинге и анализе ценных бумаг, остается элемент неточности, поскольку не всегда стоимость ценных бумаг зависит от того или иного объективного или объяснённого фактора. Но уже сейчас можно говорить о том, что МО открывает новые горизонты для анализа экономических процессов, позволяет обрабатывать большие массивы данных, выявлять скрытые закономерности, прогнозировать будущие тенденции и оптимизировать экономические решения. В 2020 г. 34% компаний в Европе, США и КНР использовали ИИ и МО, и их доля растёт [11]. На данный момент МО не автономно и не может заменить традиционные методы экономического анализа. Однако потенциально МО является прекрасным инструментом, который может использоваться в комплексе с другими методами для получения точных и достоверных результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Gogas P.* Theophilos Papadimitriou Machine Learning in Economics and Finance // Computational Economics. 2021. № 57. С. 1-4.
2. *Черкасов Д. Ю., Иванов В. В.* Машинное обучение // Наука, техника и образование. 2018. № 5. С. 85-87.
3. Machine learning for economics research: when, what and how. [Электронный ресурс]. URL: https://www.bankofcanada.ca/2023/10/staff-analytical-note-2023-16/#Introduction (дата обращения: 24.09.2024).
4. *Mullainathan S., Spiess J.* Machine Learning: An Applied Econometric Approach // Journal of Economic Perspectives. 2017. № 31. С. 87-106.
5. Fraud the facts. 2019 [Электронный ресурс]. URL: https://www.ukfinance.org.uk/system/files/Fraud%20The%20Facts%202019%20-%20FINAL%20ONLINE.pdf (дата обращения: 24.09.2024).
6. *Мабани К., Тусков А. А., Щанина Е. В.* Обнаружение мошенничества с кредитными картами с помощью машинного обучения: экспериментальный подход // Наука Красноярья. 2022. № 3. С. 17-28.
7. Credit Card Fraud Detection. [Электронный ресурс]. URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (дата обращения: 17.09.2024).
8. Credit Card Fraud Detection Predictive Models. [Электронный ресурс]. URL: https://www.kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models#Conclusions (дата обращения: 17.09.2024).
9. IBM HR Analytics Employee Attrition & Performance. [Электронный ресурс]. URL:

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (дата обращения: 17.09.2024).

10. Stock Price Analysis & Forecasting. [Электронный ресурс]. URL: https://www.kaggle.com/code/avikumart/timeseries-stock-price-analysis-forecasting/notebook#Moving-average-chart-of-'AAPL (дата обращения: 17.09.2024).

11. 17 примеров применения машинного обучения в 5 отраслях бизнеса. [Электронный ресурс]. URL: https://cloud.vk.com/blog/17-primerov-mashinnogo-obucheniya (дата обращения: 17.09.2024).