УДК 004.85:004.424

РИСКИ ИСПОЛЬЗОВАНИЯ ПРОГРАММНЫХ ПРОДУКТОВ С АЛГОРИТМАМИ МАШИННОГО ОБУЧЕНИЯ

Дедов К.К., Ковалевич А.Д.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Василькова А.Н. – ассистент кафедры ИПиЭ

Аннотация. Рассмотрены риски использования программных продуктов с использованием машинного обучения в различных сферах. Изложены наглядные примеры потенциальной некорректной работы и обучения этих программных продуктов. Определены основные причины возникновения рисков машинного обучения.

Ключевые слова: машинное обучение, искусственный интеллект, дрейф концепций, риск, программный продукт, риски ответственности.

Введение. Машинное обучение сейчас играет огромную роль, помогая организациям анализировать свои структурированные и неструктурированные данные, выявлять новые риски, автоматизировать ручные задачи в соответствии с установленными триггерами и принимать решения на основе данных. В лучшем случае он может заменить огромное количество ручного труда автоматизацией и предоставить информацию, которая поможет принять более эффективные решения в отношении оценки, мониторинга и снижения рисков. Хотя машинное обучение является инструментом управления рисками, оно также само создает множество рисков.

Основная часть. Главное отличие машинного обучения от предшествующих ему цифровых технологий заключается в способности самостоятельно принимать все более сложные решения – например, о том, какими финансовыми продуктами торговать, как автомобили реагируют на препятствия, и есть ли у пациента заболевание - и постоянно адаптироваться в ответ на новые данные. Но эти алгоритмы не всегда работают гладко. Они не всегда делают этичный или точный выбор. На это есть две фундаментальные причины.

Одна из них заключается в том, что алгоритмы обычно опираются на вероятность того, что кто-то, скажем, не выплатит кредит или заболеет. Поскольку они делают так много прогнозов, вполне вероятно, что некоторые из них будут ошибочными, просто потому, что всегда есть шанс, что они ошибутся. Вероятность ошибок зависит от множества факторов, включая количество и качество данных, используемых для обучения алгоритмов, конкретный тип выбранного метода машинного обучения. Например, глубокое обучение, которое использует сложные математические модели, в отличие от деревьев классификации, которые полагаются на правила принятия решений [1]. А также от того, использует ли система только объяснимые алгоритмы (то есть люди могут описать, как они пришли к своим решениям), что может не позволить ей достичь максимальной точности.

Вторая причина - среда, в которой работает машинное обучение, может сама развиваться или отличаться от той, для которой были разработаны алгоритмы. Хотя это может происходить разными способами, один из наиболее часто встречающихся - дрейф концепций.

Понятие дрейфа в машинном обучении и интеллектуальном анализе данных относится к изменению отношений между входными и выходными данными в основной проблеме с течением времени [2]. Рассмотрим алгоритм машинного обучения для торговли акциями. Если он был обучен на данных, полученных только в период низкой волатильности рынка и высокого экономического роста, он может оказаться неэффективным, когда экономика вступит в рецессию или испытает потрясения – например, во время кризиса, подобного пандемии вируса Ковид-19. По мере изменения рынка может меняться и связь между входом

и выходом - например, между уровнем заемных средств компании и доходностью ее акций. Подобное несоответствие может произойти и с моделями кредитного скоринга в разные моменты бизнес-цикла.

В медицине примером дрейфа концепций является ситуация, когда диагностическая система на основе машинного обучения, использующая изображения кожи в качестве исходных данных для выявления рака кожи, не может поставить правильный диагноз, потому что связь между, скажем, цветом кожи человека (который может меняться в зависимости от расы или пребывания на солнце) и решением о диагнозе не была адекватно отражена. Такая информация часто отсутствует даже в электронных медицинских картах, используемых для обучения модели машинного обучения.

Также, немаловажной проблемой являются моральные риски. Продукты и услуги, принимающие решения автономно, также должны будут решать и этические дилеммы – требование, повышающее дополнительные риски и проблемы регулирования и разработки продуктов. В настоящее время ученые начали рассматривать эти проблемы как проблемы ответственного проектирования алгоритмов. Они включают в себя загадку о том, как автоматизировать моральное мышление. Как предприятиям сбалансировать компромиссы между, скажем, конфиденциальностью, справедливостью, точностью и безопасностью? Можно ли избежать всех этих видов рисков?

Моральные риски также включают предвзятость, связанную с демографическими группами. Например, алгоритмы распознавания лиц с трудом идентифицируют людей с цветом кожи; системы классификации цвета кожи, по-видимому, имеют неодинаковую точность в зависимости от расы; инструменты прогнозирования рецидивизма дают чернокожим и латиноамериканцам ложно высокие оценки, а системы кредитного скоринга – несправедливо низкие. При широком коммерческом использовании системы машинного обучения могут быть признаны несправедливыми по отношению к определенной группе по некоторым параметрам.

Проблема усугубляется многочисленными и, возможно, взаимно несовместимыми способами определения справедливости и ее кодирования в алгоритмах. Алгоритм кредитования может быть откалиброван – это означает, что его решения не зависят от групповой принадлежности после контроля уровня риска – и при этом непропорционально часто отказывать в кредитах кредитоспособным меньшинствам. Если компания использует алгоритмы для принятия решения о том, кто получит кредит, а кто нет, ей будет трудно избежать обвинений в дискриминации некоторых групп в соответствии с одним из определений справедливости. Различные культуры могут также принимать различные определения и этические компромиссы – проблема для продуктов, имеющих глобальные рынки. Белая книга Европейской комиссии по искусственному интеллекту, опубликованная в феврале 2020 года, указывает на эти проблемы: в нем содержится призыв к разработке искусственного интеллекта с "европейскими ценностями", но будет ли он легко экспортироваться в регионы с другими ценностями?

Несовершенство машинного обучения порождает еще одну важную проблему: риски, возникающие в результате того, что программный продукт не находится под контролем конкретного предприятия или пользователя, а даже если находится, то процесс, происходящий “внутри” программного продукта, остается сокрытым от пользователя или предприятия. Такой способ работы системы называется “черным ящиком”.

Как правило, для восстановления обстоятельств, приведших к несчастному случаю, можно использовать надежные доказательства. В результате, когда такой случай происходит, руководители могут, по крайней мере, получить полезные оценки степени потенциальной ответственности своей компании. Но поскольку машинное обучение обычно встроено в сложную систему, часто бывает неясно, что привело к сбою и какая сторона (например, разработчик алгоритма или партнер) несет ответственность за ошибку. Была ли проблема с алгоритмом, с какими-то данными, переданными ему пользователем, или с данными, использованными для его обучения, которые могли быть получены от нескольких сторонних

поставщиков. Изменение окружающей среды и вероятностный характер машинного обучения еще больше затрудняют возложение ответственности на конкретного агента. Фактически, несчастные случаи или незаконные решения могут происходить даже без халатности с чьей-либо стороны – просто всегда существует вероятность принятия неточного решения. Кроме того, даже если будет установлена причина несчастного случая, виновником которого будет программный продукт, обученный с помощью алгоритмов машинного обучения, ответственность за это невозможно будет переложить на этот алгоритм. Вина будет лежать либо на разработчике программного продукта, либо на пользователе, либо разделена между всеми участниками.

Рассмотрим медицинский контекст. Суды исторически рассматривали врачей как лиц, принимающих окончательные решения, и поэтому не решались применять ответственность за качество продукции к производителям медицинского программного обеспечения. Однако ситуация может измениться по мере того, как все больше "черных ящиков" или автономных систем будут ставить диагнозы и давать рекомендации без участия (или при гораздо более слабом участии) врачей в клиниках. Такие изменения в регулировании могут переложить риски ответственности с врачей на разработчиков медицинских устройств с поддержкой машинного обучения, поставщиков данных, участвующих в разработке алгоритмов, или компании, занимающиеся установкой и внедрением алгоритмов.

Заключение. Машинное обучение представляет собой невероятно мощный инструмент, который может использоваться в самых разных областях, от финансов и банковской сферы до медицины и автомобильной промышленности. Однако, как и любая технология, машинное обучение имеет свои ограничения и недостатки, связанные с вероятностью ошибок и дрейфом концепций. Поэтому важно осознавать эти риски и использовать машинное обучение в сочетании с человеческими знаниями и опытом, чтобы достичь наилучших результатов. Только так мы сможем извлечь максимальную пользу из этой удивительной технологии, не ущемляя при этом права и интересы людей.

Список литературы

1. *Бринк, Х. Машинное обучение / Х. Бринк, Д. Ричардс , М. Феверолф – СПб.: Питер, 2018. - 336 с.*
2.*Вандер, Дж. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер – СПб.: Питер, 2018. – 576 с.*

UDC 004.85:004.424

RISKS OF USING SOFTWARE PRODUCTS WITH MACHINE LEARNING ALGORITHMS

Dedov K.K., Kovalevich A.D.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Vasilkova A.N. – assistant of the Department of EPE

Annotation. Risks of using software products using machine learning in different spheres are considered. Illustrative examples of potential incorrect operation and training of these software products are presented. The main causes of machine learning risks are identified.

Keywords: machine learning, artificial intelligence, concept drift, risk, software product, liability risks.