Организация Объединенных Наций ECE/CES/2021/11

Экономический и Социальный Совет

Distr.: General
23 March 2021
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят девятая пленарная сессия
Женева, 23–25 июня 2021 года
Пункт 5 повестки дня
Работа Группы высокого уровня по модернизации официальной статистики

Проект в области машинного обучения Группы высокого уровня по модернизации статистики

Документ подготовлен проектом в области машинного обучения Группы высокого уровня по модернизации статистики

Резюме

Настоящий документ представляет собой основной доклад проекта в области машинного обучения Группы высокого уровня по модернизации официальной статистики (ГВУ-МОС), который осуществлялся в качестве приоритетного в 2019 и 2020 годах. Цель проекта заключалась в изучении того, в каких сферах машинное обучение (МО) может способствовать повышению эффективности, а также в расширении возможностей статистических организаций по использованию машинного обучения и выявлении общих проблем в продвижении его использования в статистике.

В настоящем докладе представлена краткая информация о проделанной работе и рекомендации по продвижению использования машинного обучения в статистических организациях. Он основан на извлеченных уроках и опыте, накопленном в рамках трех рабочих модулей: пилотные исследования (РМ1), качество (РМ2) и интеграция (РМ3). Доступны отдельные доклады по трем рабочим модулям, а также сопроводительные материалы (коды, данные) размещены на вики-сайте по статистике ЕЭК ООН.

Доклад представляется Конференции европейских статистиков для информации.

I. Введение

1. Осуществление проекта в области машинного обучения (МО) было начато Группой высокого уровня ЕЭК ООН по модернизации официальной статистики (ГВУ−МОС) в марте 2019 года и завершено в декабре 2020 года. За это время в проекте приняли участие более 120 человек из 23 стран, 33 национальных и 4 международных организаций. Их совместная работа и сотрудничество были направлены на продвижение использования МО для формирования официальной статистики. Они решали эту задачу путем демонстрации полезности технологии машинного обучения для формирования официальной статистики, разработки системы качества в качестве ориентира для ее дальнейшего развития и выявления и решения проблем, связанных с интеграцией МО в производственные процессы. Доклады, документы, коды, данные и многочисленные справочные материалы были опубликованы на общедоступном вики-сайте по статистике ЕЭК ООН 13 ноября 2020 года для использования сообществом официальной статистики. После этого 16 и 17 ноября 2020 года был проведен вебинар.

2. Исходя из знаний, опыта и информации, полученных в ходе проекта, становится ясно, что машинное обучение — это не просто модный термин. Проведенные исследования показывают, что эта технология может быть интегрирована в процессы кодирования и классификации для получения более качественных результатов при тех же или более низких затратах. В некоторых случаях машинное обучение дает определенные положительные результаты в плане редактирования и импутации, однако в этой области требуются дополнительные исследования. Крайне важно использовать большие данные, например анализ спутниковых или аэроснимков. Успех машинного обучения в значительной степени зависит от объединения знаний и усилий экспертов в различных областях, в особенности в целях сбора и ведения достаточного количества высококачественных данных для эффективного обучения алгоритмов и отслеживания результатов операций, осуществляемых при содействии МО.

3. Хотя полезность МО была подтверждена пилотными исследованиями и другими недавними разработками, его интеграция в производственные процессы по-прежнему сопряжена с определенными трудностями. Проект предлагает систему качества для статистических алгоритмов и решает другие задачи в области интеграции, содействуя тем самым развитию и принятию на вооружение МО в организациях.

4. В настоящем докладе приводится справочная информация о предыстории проекта и его реализации. После перечисления основных результатов в докладе излагаются ключевые извлеченные уроки, касающиеся признания машинного обучения и содействия его продвижению, описываются соответствующие итоговые материалы проекта и предлагаются направления будущей работы.

5. На вики-сайте по статистике ЕЭК ООН доступны перечисленные ниже отдельные доклады по трем рабочим модулям, а также сопроводительные материалы (коды, данные):

a) PM1. Пилотные исследования:

- резюме пилотных исследований;
- доклады и другие документы 19 пилотных исследований, проведенных с целью оценки полезности машинного обучения при кодировании, классификации, редактировании, импутации и использовании данных изображений;
- тематические доклады, посвященные анализу подходов и результатам пилотных исследований: PM1 — Тема 1 — Доклад о кодировании и классификации; PM1 — Тема 2 — Доклад о редактировании и импутации; PM1 — Тема 3 — Доклад об анализе изображений;

b) PM2. Качество: система качества для статистических алгоритмов (СКДСА);

c) PM3. Интеграция: Доклад, посвященный выявлению и решению общих проблем, связанных с интеграцией машинного обучения в производственные процессы, на основе обследования, проведенного проектной группой;

d) другие материалы, призванные помочь пользователям в развитии машинного обучения:

- код, использованный в некоторых пилотных исследованиях — доступен в разделе «Исследования и коды»;
- два набора данных для изучения технологии машинного обучения и экспериментирования с ней — доступны в разделе «Обучение и подготовка»;
- ссылки на учебно-методические материалы — доступны в разделе «Обучение и подготовка».

II. Модернизация статистических организаций

6. Перед национальными статистическими организациями (НСО) стоит задача более оперативно реагировать на растущие потребности в более актуальной, своевременной, подробной и доступной статистической информации и сервисах данных, которым можно доверять и которые можно использовать для принятия ясных и эффективных политических решений. Кроме того, НСО требуется подтверждать свою высокую эффективность и оправдывать ожидания в существующих бюджетных рамках. Они также сталкиваются с вызовами, обусловленными постоянно растущим объемом данных, доступных в самых разных источниках и форматах и с разным уровнем качества.

7. И наконец, НСО должны конкурировать со все большим числом государственных и частных организаций, больших и малых, занимающихся формированием и распространением статистики более своевременными и доступными способами, что привлекает внимание политиков и многих других пользователей, зачастую несмотря на недостатки в плане актуальности или качества. Это происходит по ряду причин, в частности благодаря разработке этими организациями альтернативных методов сбора данных, например алгоритмов машинного обучения, или наличию у них оперативного доступа к этим данным и возможности включать их в производственный процесс, наличию более мощного ИТ-потенциала и установлению ими меньшего количества ограничений в плане качества, прозрачности, этики и неприкосновенности частной жизни.

8. В то же время именно в этих областях НСО обладают определенным конкурентным преимуществом. Публикуя подробную информацию об источниках данных, методах и различных показателях, НСО демонстрируют высокий уровень прозрачности. Они обладают значительным коллективным опытом эффективной интеграции различных источников данных и имеют юридическое обязательство по уважению неприкосновенности частной жизни и обеспечению защиты от разглашения. Кроме того, они способны еще больше укрепить свой потенциал в этих областях, задействуя сети специалистов по всему миру, которые совместно работают над удовлетворением общих потребностей и решением приоритетных задач.

9. Помимо использования своего индивидуального и коллективного опыта, статистические организации должны обладать способностью к адаптации, чтобы сохранять свою актуальность, своевременно и непрерывно реагируя на потребности заинтересованных сторон в статистических данных и услугах. Одним из краеугольных камней, позволяющих НСО выполнять свой мандат, является разработка надежных методов и процессов, интегрированных в процесс формирования официальной статистики. Как отмечалось выше, это осложняется ростом спроса на данные и увеличением количества источников, а также производителей информации, возможности которых все легче увязывать источники данных с поступающими запросами постоянно расширяются.

III. Предложение по проекту машинного обучения

10. Вышеупомянутые проблемы находятся в центре внимания Группы высокого уровня ЕЭК ООН по модернизации официальной статистики (ГВУ-МОС) — группы преданных своему делу главных статистиков, активно руководящих процессом модернизации национальных статистических организаций. Их миссия заключается в совместной работе по выявлению тенденций, вызовов и возможностей в области модернизации этих организаций. ГВУ-МОС обеспечивает общую платформу для экспертов в целях разработки решений гибким и оперативным образом.

11. ГВУ-МОС опирается на Исполнительный совет, отвечающий за стратегическое управление текущей деятельностью, включая работу четырех постоянных рабочих групп и осуществление двух проектов с ограниченным сроком реализации. Одна из рабочих групп, Сеть передовых исследований и изысканий (СПИИ), представляет собой перспективную «фабрику идей» для сообщества модернизации статистики. Члены СПИИ признают, что многие новые производители данных используют широкое разнообразие источников, применяя подходы и методы, отличающиеся от тех, которые традиционно используются статистическими организациями. Многие из этих источников данных требуют инновационных подходов и методов, таких как машинное обучение и искусственный интеллект. На своем рабочем совещании в ноябре 2018 года ГВУ-МОС еще раз признала важность интеграции этих технологий в процесс формирования официальной статистики, поддержав предложение СПИИ о запуске проекта в области машинного обучения. Подготовленный СПИИ программный документ содержал следующее обоснование проекта:

> «Интерес к использованию машинного обучения для целей официальной статистики быстро растет. Для обработки некоторых вторичных источников данных (включая административные источники, большие данные и Интернет вещей) представляется необходимым изучить возможности, предоставляемые современными методами МО, поскольку они могут оказаться полезны и для первичных данных, о чем свидетельствует программный документ по МО. Хотя технология МО выглядит многообещающей, статистическое сообщество ЕЭК ООН обладает лишь ограниченным опытом в его использования в конкретных добавлениях, и некоторые вопросы, связанные с качеством и прозрачностью результатов, полученных с помощью МО, по-прежнему требуют решения».

IV. О проекте машинного обучения

12. В конце февраля 2019 года ЕЭК ООН произвела найм руководителя проекта, а в марте приступила к его осуществлению совместно с 11 участниками из 6 организаций. Были проведены виртуальные совещания для определения сферы охвата и планирования проекта. Эти планы были окончательно доработаны на первом очном совещании, организованном Управлением национальной статистики Соединенного Королевства в мае 2019 года. К тому времени в проекте были задействованы уже 27 участников из 14 организаций. Они достигли согласия по перечисленным ниже ключевым аспектам.

13. На основе взаимных интересов и существующих научных разработок было определено, что цель проекта заключается в продвижении исследований, разработке и применении методов машинного обучения для повышения эффективности процесса формирования официальной статистики. Для достижения этой цели группа проекта машинного обучения будет стремиться:

- изучить и продемонстрировать полезность МО для формирования официальной статистики, причем под «полезностью» понимается повышение актуальности, улучшение общего качества или сокращение затрат;
- расширять возможности использования МО для повышения эффективности процесса формирования официальной статистики;

- укреплять потенциал национальных статистических организаций по использованию МО для формирования официальной статистики;
- укреплять сотрудничество между статистическими организациями в деле разработки и применения МО.

14. Исходя из первых итогов работы СПИИ, ожиданий ГВУ-МОС и интересов членов группы, работа проекта была организована в соответствии с тремя рабочими модулями (РМ):

- РМ1— проведение пилотных исследований по следующим темам:
  - кодирование и классификация (К&К);
  - редактирование и импутация (Р&И);
  - использование изображений;
- РМ2 — разработка системы качества, подводящей фундамент под использование МО;
- РМ3 — выявление и решение проблем, связанных с интеграцией.

15. Помимо демонстрации полезности МО, целью пилотных исследований являлось развитие обучения, обмена информацией и сотрудничества в рамках группы. В начале проекта отсутствие системы качества, призванной служить ориентиром для использования МО, рассматривалось в качестве главного препятствия появлению в рамках НСО доказавших свою эффективность решений МО. И наконец, первоначально были обозначены и другие проблемы, связанные с внедрением и дальнейшим распространением решений МО, и по мере продвижения пилотных исследований их число увеличивалось.

16. Если свести цели проекта и рабочие модули в одно предложение, то проект был начат с целью «интеграции в производственные процессы (РМ3) демонстрируемых решений МО (РМ1) рациональным и эффективным образом (РМ2)». Участники из Германии, Канады, Мексики и Соединенного Королевства выразили готовность возглавить пять подгрупп (три подгруппы по РМ1; одна подгруппа по РМ2; одна подгруппа по РМ3). Подгруппы проводили виртуальные совещания в целях совместной работы и обмена информацией о достигнутом прогрессе и результатах. Совещания, посвященные ходу осуществления проекта, проводились ежемесячно.

17. По мере развития проекта к нему присоединялось все больше участников. В нем также принимали участие и другие люди, работавшие над пилотными исследованиями (коллабораторы), и люди, проявившие интерес к отслеживанию хода осуществления проекта (заинтересованные лица). На момент написания настоящего доклада в проекте участвовали 124 представителя 23 стран, 33 национальных и 4 международных организаций. Как следствие, на ежемесячных совещаниях регулярно присутствовали от 40 до 60 человек. Хотя проект был сосредоточен на машинном обучении, благодаря его широкому членскому составу к его реализации были привлечены эксперты из других международных групп, работающих над темами, охватывающими вопросы машинного обучения, или проявляющих интерес к ним, например Глобальной группы Организации Объединенных Наций по большим данным и Группы ГВУ-МОС ЕЭК ООН по развитию потенциала и информационному взаимодействию.

18. В целях дальнейшего развития обмена информацией и сотрудничества было проведено четыре совещания по проекту:

- в мае 2019 года состоялось совещание, организованное УНС Соединенного Королевства и посвященное окончательной доработке целей проекта, его организации и первоначальных результатов;
- в сентябре 2019 года состоялось совещание, организованное Статистическим управлением Республики Сербия для обмена информацией о достигнутом прогрессе, укрепления механизмов сотрудничества и согласования результатов и сроков осуществления проекта;

- в апреле 2020 года состоялось совещание, проведенное в режиме онлайн из-за пандемии (первоначально оно должно было быть организовано Статистическим управлением Польши) для обмена информацией о результатах пилотного исследования и их обсуждения;
- в октябре 2020 года состоялось совещание, также проведенное в режиме онлайн, для обмена информацией и обсуждения докладов пилотных исследований, тематических докладов пилотных исследований, итогов проекта и будущих направлений деятельности. На этом совещании участникам проекта было также предложено рассказать о других сферах применения машинного обучения в их организациях.

19. Все документы по проекту размещались для ознакомления участников проекта на вики-странице по статистике ЕЭК ООН. Они включали рабочие документы, информацию о ежемесячных совещаниях и мероприятиях, а также множество справочных материалов, представленных участниками проекта. Эти документы также включали в себя обсуждения по проекту и его итоговые материалы, представленные участниками проведенных мероприятий, например конференции BigSurv20 («Большие данные и научные исследования») и Geo Week 2019 (Группа по наблюдениям Земли, Саммит 2019 года), и доклады, представленные на ежегодном рабочем совещании ГВУ−МОС по модернизации. Наконец, участники использовали различные средства для распространения кода МО (в основном через GitHub) и некоторых других данных для облегчения и ускорения обучения и экспериментов других участников.

20. Многие из документов и других материалов, упомянутых выше, были опубликованы на вики-странице по официальной статистике ЕЭК ООН 13 ноября 2020 года. После этого 16 и 17 ноября 2020 года был проведен вебинар, на котором были представлены итоги проекта, а в заключение состоялось открытое обсуждение будущих направлений работы. В вебинаре приняли участие 203 человека из 33 стран и 60 национальных и международных организаций.

21. Конечным итогом проекта в области МО стало формирование большой группы экспертов в различных областях и представителей организаций, занимающихся продвижением использования МО, которые готовы совместно работать в этом направлении. Научно-исследовательский центр интеллектуальной обработки данных Управления национальной статистики Соединенного Королевства возглавил группу МО 2021 при поддержке секретариата ЕЭК ООН и Исполнительного совета ГВУ-МОС.

V. Итоги проекта, извлеченные уроки и другие вопросы

22. В рамках трех рабочих модулей было проведено 21 пилотное исследование[1], разработана система качества и подготовлен доклад, в котором были выявлены и рассмотрены общие проблемы, связанные с интеграцией МО в производственные процессы. Информация о многих уроках, извлеченных в ходе этой работы, была размещена на вики-странице ЕЭК ООН (см.: «Проект ГВУ-МОС в области машинного обучения»). Результаты и извлеченные уроки каждого пилотного исследования (см.: доклады пилотных исследований) были проанализированы и обобщены в докладах по каждой из трех областей их проведения (см.: «Доклад о кодировании и классификации»; «Доклад о редактировании и импутации»; «Доклад об анализе изображений»). Последние доклады были дополнительно проанализированы и обобщены, с тем чтобы выделить основные результаты и извлеченные уроки в области применения МО в процессах формирования статистики (см. резюме PM1). Деятельность в рамках рабочих модулей по качеству и интеграции затрагивала

[1] К числу 21 проведенного исследования относится создание общей структуры для использования МО в целях анализа изображений и разработка документа, содержащего «идеи и подсказки» по применению МО для редактирования данных. Эти итоги не основаны на каком-либо конкретном примере использования МО.

ключевые аспекты, необходимые для содействия продвижению и признанию МО (см.: «Система качества для статистических алгоритмов»; «Доклад об интеграции»).

23. Настоящий доклад призван информировать читателя об извлеченных уроках и других соображениях, которые высказывались на протяжении всего проекта, избегая при этом, насколько это возможно, повторения выводов других докладов. Группа пришла к выводам о том, что:

- проект достиг своих целей;
- большинство участников приобрели обширные знания и опыт в использовании МО;
- проект способствовал продвижению использования МО в целях формирования официальной статистики в большинстве участвующих организаций;
- однако имеется недостаточно доказательств успешного применения добавлений МО в производственных процессах;
- существует больший интерес и потребность в том, чтобы превратить МО из многообещающего в реальное производственное решение.

24. Основываясь на знаниях и опыте, полученных в ходе осуществления проекта, продвижение МО в целях формирования официальной статистики можно обобщить в двух словах: принятие к использованию и содействие. Большая часть нагрузки по принятию к использованию МО в статистической организации ложится на лиц, разрабатывающих и внедряющих соответствующие методы. Основная задача по содействию развитию и практическому применению МО лежит на организации. Важнее всего то, что как для принятия к использованию МО, так и для содействия его применению требуется поддержка всех сотрудников.

25. В ниже следующих разделах рассматриваются ключевые аспекты принятия к использованию решений МО и содействия их применению. Каждый аспект кратко обсуждается, увязывается с результатами, достигнутыми в ходе осуществления проекта, и завершается предложениями по будущей работе для группы МО 2021. В приводимой в добавлении таблице 1 эти ключевые аспекты изложены в порядке от формулирования идеи до производственного сопровождения. В ней также приводится информация об основополагающих элементах и поддержке со стороны организаций. Таблица иллюстрирует вопросы, рассмотренные в рамках проекта по МО и некоторые аспекты, определенные в качестве потенциальных тем для изучения группой МО 2021.

VI. Ключевые аспекты принятия к использованию решений машинного обучения

A. Увязка с бизнес-потребностями

26. Решения МО должны быть в конечном итоге приняты к использованию людьми, ответственными за формирование данных (как правило, занимающимися этой темой аналитиками), и, что более важно, теми, кто пользуется данными. Как и любой подход или технология, МО является одним из средств достижения определенной цели. Его следует рассматривать и применять не с точки зрения его содержания, а с точки зрения того, как оно может способствовать более эффективному удовлетворению бизнес-потребностей (повышению актуальности, детализации, своевременности, точности, эффективности затрат и т. д.). Пилотные исследования в целом были сосредоточены на вопросах повышения своевременности и точности трех статистических процессов (см. доклады пилотных исследований). Существует множество примеров применения МО для удовлетворения прочих бизнес-потребностей в рамках других процессов (некоторые примеры см.: «Другие сферы применения машинного обучения»).

27. Одной из отличительных черт проекта, способствовавших его эффективному осуществлению, являлся его подход практического эксперимента. На ранней стадии

проекта возникла идея создать своеобразную «книгу рецептов» по использованию МО. В рамках проекта были разработаны некоторые основные элементы правильного «рецепта» (рамки и передовой опыт), в то время как многие участники обучались, экспериментируя с различными «ингредиентами» (пилотные исследования), для удовлетворения потребностей своих организаций. В дальнейшем важно, чтобы деятельность группы по-прежнему основывалась на потребностях участвующих организаций (применение) и вносила вклад в более фундаментальные аспекты (см. строку «Поддержка» в таблице 1 в добавлении).

B. Руководящие принципы системы качества

28. Решения МО должны способствовать достижению качественных или еще более качественных результатов для удовлетворения бизнес-потребностей. Для этого необходимо определить, что представляет собой «качество». Его определения содержатся во многих общепринятых системах качества, разработанных национальными и международными статистическими организациями. Система качества для статистических алгоритмов (СКДСА) дополняет эти системы, фокусируясь на аспектах, которые наиболее важны для принятия к использованию решений МО (см.: «Система качества для статистических алгоритмов»). СКДСА содержит рекомендации по выбору алгоритмов (в том числе традиционных) для производственного процесса. В ней целенаправленно используется терминология статистических алгоритмов, поскольку она охватывает как традиционные, так и современные методы, обычно используемые официальными статистическими органами, для укрепления взаимопонимания между сторонниками каждого из них. Не существует какой-либо устоявшейся формулы, позволяющей констатировать, что решения МО дают такие же или лучшие результаты по сравнению с альтернативными решениями. Как и большинство систем качества, СКДСА содержит пять элементов, которые должны рассматриваться в совокупности. Можно уделять большее внимание одному или двум из них, но ни один из них не должен игнорироваться.

29. СКДСА разрабатывалась одновременно с проведением пилотных исследований. Эти исследования внесли важный вклад в работу и способствовали развитию системы, однако времени для формального экспериментального применения отдельных видов практики, рекомендованных в соответствии с ней, было недостаточно. Одной из отличительных черт проекта, как было отмечено некоторыми его участниками, был его подход практического эксперимента. В дальнейшем рекомендуется проводить такие эксперименты в рамках пилотных исследований в целях получения ценной обратной связи для совершенствования и расширения системы качества.

C. Демонстрация полезности

30. Этому аспекту было посвящено большинство пилотных исследований. Что касается кодирования и классификации, то исследования показали, что результаты при использовании МО могут быть лучше, чем при проведении операций исключительно вручную. Общая проблема, отмеченная в ходе пилотных исследований, заключалась в отсутствии статистически достоверного исходного показателя, с которым можно было бы сравнить результаты использования МО. Как следствие, в начале многих исследований ставилась цель по тиражированию уже применяемой операции, например создания тех же категорий продуктов, что и при ручной классификации, и повышению эффективности, главным образом в том, что касается своевременности и, опосредованно, затрат. В этой связи возникают три серьезные проблемы. Во-первых, точность применяемой (или альтернативной) операции часто либо неизвестна, либо не подтверждается надежным методом оценки. Во-вторых, МО никогда не сможет полностью повторить другую операцию. Исследования, проведенные в рамках проекта, показывают, что в результате применения МО можно с высокой степенью точности воспроизводить от 40 % до 85 % результатов применяемой (ручной или другой автоматизированной) операции. Однако, в-третьих,

и это более важно, цель применения МО не должна ограничиваться повторением другой операции, если только в результате она не будет проведена намного быстрее и при значительно меньших затратах. Цель должна заключаться в улучшении операции путем объединения соответствующих сильных сторон каждого из методов. В контексте операции классификации это может означать использование расчетных данных МО для автоматического присвоения категории (на основе расчетов, подтвердивших свою высокую точность, например более 98 %); использование недостаточно точных расчетов для поддержки работы кодировщиков; и отказ от неточных расчетов и использование кодировщиков для классификации остальных (часто менее распространенных) категорий. В производственном процессе используются различные варианты этой стратегии (см.: «Производственные травмы и заболевания»; «Отрасли и виды занятий»; «Стандартная отраслевая классификация»). Был также проведен эксперимент по применению этой стратегии с использованием общего кода МО и данных (см.: «Опыт пользователя по применению общего кода МО и данных»; «Общий код»; «Набор данных для описания продуктов»).

31. Что касается редактирования и импутации, то исследования показали различные результаты, начиная от полного отсутствия полезности МО (простая импутация была более эффективна, чем другие методы) и заканчивая наличием перспектив. Нет никаких свидетельств того, что методы МО не работают. Для них может потребоваться меньше программирования и меньше времени для внедрения, по сравнению с используемыми методами. С другой стороны, получение и хранение качественных данных для обучения таких алгоритмов является сложной задачей, равно как и объяснение того, какого результата можно добиться за счет применения МО, и того, как именно это происходит, для привлечения заинтересованных сторон, даже если предлагаемый процесс оказывается более быстрым и точным. Для выработки руководящих принципов применения МО в этой области и определения благоприятных условий для его использования необходимы дополнительные исследования и фундаментальные разработки (такие, как «Советы и идеи по очистке данных»).

32. С самого начала проекта считалось, что МО необходимо для эффективного использования больших массивов данных. Это было подтверждено в ходе пилотных исследований по анализу изображений (спутниковая и аэросъемка). По мере расширения доступа к большим массивам таких данных одной из задач становится предоставление пользователям информации о сложных процессах, необходимых для их правильного и эффективного использования, в том числе в тех случаях, когда требуется машинное обучение. Для предоставления части этой информации в ходе проекта было предложено создать типовой процесс формирования официальной статистики с использованием спутниковых данных и машинного обучения (см.: «Типовой процесс»). Он использовался для описания двух исследований по спутниковым и аэроснимкам (см.: «Классификация использования адресов аэроснимков» и «Интеграция НЗ в официальную статистику с использованием машинного обучения»).

33. В дальнейшем организациям рекомендуется продолжать развитие своих текущих разработок в области МО в направлении практического внедрения, продолжая при этом сотрудничество и обмен опытом с другими субъектами. Эта деятельность может быть расширена, с тем чтобы охватить другие области, представляющие интерес для организаций (см. некоторые примеры: «Другие сферы применения машинного обучения»), в частности те бизнес-потребности, которые требуют значительных трудозатрат, не меняются с течением времени и предлагают большие объемы данных для обучения алгоритмов. В рамках этой деятельности следует рассмотреть возможность экспериментального применения некоторых методов, предлагаемых в СКДСА и документах по интеграции, и предоставления ценной обратной связи на основе накопленного опыта.

D. Эффективность функционирования во времени

34. Пилотные исследования были сосредоточены на оценке полезности различных алгоритмов МО и определении наилучшей модели (алгоритма и параметров) на основе имевшихся данных. Как уже было отмечалось ранее, по-прежнему существует множество проблем, связанных с внедрением демонстрируемого решения МО в производственный процесс. Не менее важно и то, что такое решение МО с течением времени должно не только сохранять свою эффективность, но и повышать ее, так как системы МО «учатся» и адаптируются по мере эволюции вводных данных. В ходе рассмотрения решений МО в рамках проекта участники интересовались, когда, как часто и как именно следует обновлять или корректировать алгоритмы МО и/или его параметры. Аналогичные вопросы возникали и в связи с другими сферами применения МО.

35. Только одно из приложений МО, рассматривавшихся в рамках проекта, использовалось в производственном процессе достаточно долго для того, чтобы накопить значительный опыт по обновлению его алгоритмов (см.: «Производственные травмы и заболевания»). Центральным элементом разработки и сопровождения эффективного решения МО являются данные, используемые для обучения, не только на начальном этапе — при настройке исходного алгоритма и его параметров — но и на протяжении всего процесса его использования. Еще одним ключевым элементом являются данные, используемые для оценки, иногда называемые «золотым стандартом». Эти данные необходимы для оценки не только того, как работает алгоритм МО, но и всей операции, так как обычно она включает в себя некоторые конторские операции. Они должны быть независимы от данных, используемых для обучения. Эти данные имеют важное значение, но, как правило, требуют значительных затрат и должны соответствовать определенным требованиям, например по сбору данных подспутниковых наблюдений или классифицированию текстов экспертами в соответствующей области. В ходе проекта удалось оценить важность и проанализировать характеристики качественных данных для обучения и оценки.

36. В дальнейшем рекомендуется задокументировать и распространить эти знания.

E. Соблюдение этических и правовых норм

37. «На протяжении последних десяти лет популярность машинного обучения растет, а областей его применения становится все больше. В частности, оно применяется в социальной сфере, где составление моделей, основанных на данных профилирования, может оказывать значительное влияние на жизнь людей. Для предотвращения нежелательной дискриминации в этих моделях были предложены различные методы для обеспечения объективности алгоритмов», — этот текст взят из аннотации к рабочему докладу Центра по статистике больших данных Статистического управления Нидерландов (см.: «Объективные алгоритмы в контексте»). Эта цитата подчеркивает важность вопросов этики, которые были затронуты в ходе нескольких обсуждений в рамках проекта, но не рассматривались отдельно.

38. В дальнейшем эти вопросы могли бы быть изучены либо группой по МО, либо в сотрудничестве с другими рабочими группами, которые могли бы рассмотреть их в более широком контексте. В этой работе важно проводить различие между проблемами, связанными с источниками данных, и методами их использования. Кроме того, важно сосредоточить внимание на аспектах официальной статистики, которые, как часто упоминалось в обсуждениях, в основном представляют собой обобщенные данные, а не конкретные результаты по отдельным лицам, например в связи с одобрением кредитов или постановкой медицинского диагноза.

F. Ведение разработок на прочной научной и многодисциплинарной основе

39. Национальные и международные официальные статистические организации всегда формируют актуальную и достоверную информацию, поскольку их деятельность основывается на надежных методах и процессах. Когда методы МО разрабатываются и внедряются на тех же базовых принципах, они во многом способствуют решению вышеперечисленных проблем и поэтому охотно принимаются к использованию. В основе этих процессов должна лежать научная база, охватывающая знания и навыки по многим дисциплинам: тематическим дисциплинам, статистике, информатике, методологии, обработке данных и операциях с ними. По сравнению с традиционными методами, эти дисциплины должны быть еще более тесно увязаны друг с другом с самого начала процесса (возникновение идеи и ее увязка с бизнес-потребностью) до его завершения (практическое применение). Это особенно актуально для тематических знаний, в случае которых МО является не просто еще одним решением для удовлетворения конкретных бизнеса-потребностей, но и само нуждается в таких знаниях для своего функционирования. В то время как идея использования МО может исходить от одного человека (как в некоторых пилотных исследованиях), для ее дальнейшего эффективного развития необходимо оперативное подключение других дисциплин, в особенности тематических, для правильного и эффективного продвижения вперед. Как и в случае с аспектом качества, можно выделить одну или две конкретных дисциплины, но ни одна из них не должна быть упущена из виду. Это было подчеркнуто в ходе эксперимента с кодом МО и обменом данными в рамках проекта. Этот эксперимент проводился человеком, имевшим ограниченные знания о МО, который по ходу его осуществления многому научился и совершил много ошибок (см.: «Опыт пользователя по применению кода МО и обмену данными»).

40. Проект МО выиграл от привлечения экспертов из многих областей. Это позволило изучить и распространить различные точки зрения и аспекты, которые следует учитывать при разработке, оценке и продвижении решений в области МО. В дальнейшем, в зависимости от своих планов и целей, группе по МО следует привлекать больше участников, имеющих отношение к обработке данных, тематическим дисциплинам и информационным технологиям. Статистические организации будут продолжать сталкиваться с проблемой получения, развития и упорядочения разнообразного опыта, необходимого для эффективного и результативного использования МО для удовлетворения бизнеса-потребностей. Получение и развитие (например, за счет обучения) экспертных знаний было названо в качестве наиболее насущной потребности в ходе опроса, проведенного на вебинаре (см.: вебинар по МО; результаты опроса). Этот аспект более подробно рассматривается ниже в контексте содействия принятию к использованию решений МО.

VII. Ключевые аспекты содействия принятию к использованию решений МО

A. Комбинирование междисциплинарных навыков

41. Процесс формирования официальной статистики всегда комбинировал и продолжает комбинировать знания и экспертизу многочисленных дисциплин. Это применимо и в данном случае, тем более в условиях увеличения количества источников данных (больших, средних или малых), пользователей, стремящихся их использовать, и технологий, позволяющих их использование. Хотя многие из этих навыков используются в интеллектуальной обработке данных (относительно новая дисциплина), масштабы и глубина знаний, необходимых в каждой из дисциплин, не могут быть получены от одного или нескольких человек. Сведение воедино необходимых навыков является одной из основных задач, стоящих перед статистическими организациями. Эту задачу можно разделить на четыре подзадачи:

выявление, приобретение, развитие и организация. Некоторые из них были рассмотрены в рамках проекта (см.: «Интеграция»). При этом было определено множество конкретных мер для НСО, которые могли бы содействовать использованию МО и способствовать его расширению. Многие из них были разработаны совсем недавно (см.: «Инициативы по ускорению интеграции решений машинного обучения»).

42. Эти инициативы включают в себя создание отдельных организаций, занимающихся интеллектуальной обработкой данных, лабораторий и внутренних или внешних форумов для обмена информацией по интеллектуальной обработке данных и машинному обучению. Руководители этих организаций будут скорее всего находиться в контакте и обмениваться информацией на неофициальной основе. В дальнейшем им рекомендуется создать официальную сеть для обмена информацией о существующих проблемах, практике, полученном опыте и результатах. Работа этой сети должна быть сосредоточена на управленческих аспектах, например на корпоративных стратегиях, удовлетворении потребностей, изменении культуры, коммуникации. Сеть будет также тесно взаимодействовать с группой МО 2021, например предлагать проекты и определять их приоритетность. Сеть или проект МО 2021 также должны быть связаны с другими группами, работающими над аналогичными вопросами, например с Группой ГВУ-МОС ЕЭК ООН по развитию потенциала и информационному взаимодействию, работающей, помимо прочего, над управлением изменениями, разработкой организационных рамок для сотрудничества и формированием компетенций.

B. Вычислительная инфраструктура

43. С самого начала проекта по МО было решено сосредоточить внимание на демонстрации полезности, вопросах качества и интеграции. Вопрос вычислительной инфраструктуры обсуждался в ходе его осуществления, но лишь в самых общих чертах. В дальнейшем его изучение следует продолжить в рамках проектов МО 2021, однако до этого необходимо понять, существуют ли другие рабочие группы или другие инициативы по этому вопросу, и установить с ними связь.

C. Исследования и разработки

44. Первым ключевым аспектом принятия к использованию вышеупомянутых решений МО является их увязка с бизнес-потребностями. В ходе обсуждений в подгруппе проекта по интеграции не был достигнут полный консенсус по этому вопросу. Некоторые подчеркивали необходимость начинать с бизнес-потребностей, затем переходя к исследованиям и разработкам, созданию прототипа и внедрению других аспектов, таких как информационные технологии. Другие подчеркивали важность накопления опыта в области МО, в первую очередь, посредством НИОКР, что, в свою очередь, позволит выявить соответствующие бизнес-проблемы, которые могут быть решены с помощью машинного обучения. В дальнейшем какой бы путь развития МО не был выбран, он должен определяться если не конкретными бизнес-потребностями (например, одной статистической программы), то, по крайней мере, четкой общеорганизационной стратегией для постоянного повышения актуальности за счет предоставления доступа к большему объему информации лучшего качества в более сжатые сроки и с потенциально более низкими затратами.

D. Обмен информацией и сотрудничество

45. В рамках проекта его участники обменивались рабочими документами, методологическими и техническими справочными материалами, ссылками на учебные ресурсы, презентациями на встречах и спринтах. Доступ к экспертным ресурсам и возможностям для проведения совещаний часто обеспечивался через вспомогательную структуру ГВУ-МОС. Обмен кодом МО и данными значительно облегчил и ускорил обучение и эксперименты других участников. Многие из

документов и других материалов, упомянутых выше, были сгруппированы и опубликованы на общедоступной вики-странице ЕЭК ООН 13 ноября, чтобы любой член сообщества официальной статистики мог воспользоваться знаниями и материалами, накопленными проектной группой за два года (см.: доклады о пилотных исследованиях; «Общий код»; «Обучение, справочные материалы и обмен данными»).

46. В дальнейшем обмен информацией и сотрудничество не только пойдут на пользу развитию МО, но и позволят быстро сориентироваться и избежать его применения в тех областях или контекстах, где, как было установлено, оно не приносит никакой пользы. Несмотря на то, что виртуальные спринты и другие совещания были успешными, учитывая сложившиеся обстоятельства, в будущем будет важно периодически проводить очные совещания в форме семинаров экспертов или рабочих совещаний, в ходе которых может быть налажено более тесное сотрудничество и которые могут дополнительно подкрепляться последующим дистанционным обменом мнениями. Вспомогательная структура ГВУ-МОС должна и далее способствовать развитию МО путем предоставления доступа к информации и ее распространения среди участников широкой сети заинтересованных сторон.

E. Поддержка со стороны старшего руководства

47. Проект в области машинного обучения не существовал бы и не был бы успешным без участия множества людей из многочисленных организаций и поддержки со стороны главных статистиков через ГВУ-МОС. В дальнейшем их поддержка будет жизненно необходима для проведения исследований и разработок в их соответствующих организациях и вместе с другими участниками в рамках совместных инициатив. В свою очередь, эти инициативы должны соответствовать приоритетам статистических организаций. В случае МО 2021 это рекомендуется делать, продолжая представлять доклады ГВУ-МОС через ее Исполнительный совет.

F. Вовлечение всех сотрудников

48. Новые технологии, такие как машинное обучение или искусственный интеллект, оказывают значительное влияние на культуру организации. МО способно изменить то, что может сделать организация и каждый отдельный сотрудник, и как они могут это сделать. Во всех исследованиях, проведенных в рамках проекта, использовались контролируемые методы обучения, которые требуют существенного вовлечения всех сотрудников, в частности сотрудников соответствующих тематических и технических отделов, которые с наибольшей вероятностью будут затронуты изменениями. Проведенные исследования демонстрируют, что методы МО не могут полностью заменить труд персонала и должны восприниматься не в качестве таковых, а скорее как средство внедрения или укрепления автоматизированных процессов для достижения более высоких результатов при тех же или более низких затратах и предоставления сотрудникам времени для выполнения работы более эффективным образом. Поскольку решения МО доказали свою полезность во многих операциях, необходимо поощрять сотрудников на всех должностях и на всех уровнях организации к тому, чтобы рассматривать МО в качестве потенциального решения для удовлетворения своих бизнес-потребностей. Они также должны иметь доступ к экспертам или центрам экспертных знаний по МО, чтобы оперативно определять, следует ли продолжать рассматривать вопрос об использовании МО.

VIII. Вывод: машинное обучение — мода, необходимость или пустышка?

49. В своем программном документе, подготовленном в декабре 2018 года и содержащем предложение по осуществлению проекта в области машинного обучения, члены Сети передовых исследований и изысканий написали: «Хотя технология МО выглядит многообещающей, статистическое сообщество ЕЭК ООН имеет лишь ограниченный опыт его практического применения, и некоторые вопросы, связанные с качеством и прозрачностью результатов применения МО, по-прежнему требуют

решения». В то время эту фразу в контексте формирования официальной статистики можно было бы сократить до следующего вопроса: «Машинное обучение — мода, необходимость или пустышка?».

50. После двух лет работы над проектом в области машинного обучения можно сделать вывод, что МО — не просто модная концепция; оно абсолютно необходимо в тех сферах, где оно может способствовать повышению эффективности, и не должно использоваться в остальных (не приводя, таким образом, к провалам); и что по-прежнему сохраняются отдельные проблемы, которые признаются и решаются некоторыми участниками процесса.

51. Об интересе к МО и прогрессе в его использовании свидетельствует постоянный рост числа участников проекта, который в настоящее время насчитывает 124 члена из 23 стран, 33 национальных и 4 международных организаций. Активное участие многих из них при поддержке других позволило подготовить и распространить в рамках проекта множество докладов, документов, кодов, данных, учебных материалов и других справочных документов, которые могут быть полезны для сообщества официальной статистики. Помимо проекта, стремительно растет число лекций, сессий или целых мероприятий, посвященных машинному обучению и интеллектуальной обработке данных (см. список сессий и лекций на BigSurv20 — «Большие данные и научные исследования», в разделе «Другие сферы применения машинного обучения»; следующий симпозиум Статистического управления Канады будет посвящен интеллектуальной обработке данных)[2].

52. Машинное обучение абсолютно необходимо в тех сферах, где была доказана его эффективность в подготовке более актуальных, качественных, своевременных и выгодных с точки зрения затрат данных, без значительного ослабления любого из этих аспектов. Оно с большей вероятностью даст лучшие результаты в трудоемких, повторяющихся и стабильных операциях, таких как кодирование и классификация. Можно сказать, что это происходит при любой автоматизации таких операций. В случае МО эта автоматизация может быть выполнена быстрее. МО необходимо во многих случаях использования больших массивов данных. Сложнее обстоит дело с процессами, подразумевающими высокую степень субъективности, например с редактированием и импутацией.

53. Сталкиваясь со все новыми и постоянно эволюционирующими технологиями, некоторые стороны пытаются им сопротивляться. Одни выступают против них, приводя научные аргументы, которые только их укрепляют. Другие просто противятся им, как большинству изменений. Первых можно будет убедить, если решения МО будут разрабатываться на прочной научной и многодисциплинарной основе, в которой они нуждаются, и соответствовать требованиям системы качества и этическим соображениям. В случае последних поможет четкая и сильная поддержка со стороны старшего руководства. Обмен информацией и сотрудничество внутри статистических организаций и между ними также важны для продвижения использования МО на основе извлеченных уроков, благодаря которым было установлено, в каких сферах оно будет полезно, в каких оно демонстрирует многообещающие результаты, а в каких его не следует применять.

54. За полтора года проект в области машинного обучения придал значительный импульс продвижению идеи ответственного использования МО для формирования официальной статистики. Проект способствовал продвижению использования МО в участвующих организациях и обмену многочисленными извлеченными уроками с сообществом официальной статистики. Не менее важно то, что итогом проекта стало формирование сильной группы экспертов из различных областей, желающих продолжать работу, наличие организации, готовой взять на себя руководящую роль и заинтересованность других людей в присоединении к группе и ее поддержке на пути к продвижению решений МО на протяжении всего процесса от выдвижения идеи до ее практического применения.

[2] Ссылка на его анонс будет доступна в ближайшее время.

Таблица

Рабочие модули, исследованные в рамках проекта МО 2020 (гиперссылки), и потенциальные темы, которые будут изучены в рамках проекта МО 2021 (красный текст)

<table>
<tr><td rowspan="5">Процесс</td><td colspan="3">Переход от идеи к работоспособному решению (демонстрация)</td><td colspan="2">Переход от работоспособного решения к производству (практическое применение)</td><td colspan="2">Обеспечение эффективности функционирования (сопровождение)</td></tr>
<tr><td colspan="3">Все пилотные исследования РМ1</td><td colspan="2">Некоторые пилотные исследования РМ1</td><td colspan="2">Лишь немногие пилотные исследования РМ1</td></tr>
<tr><td colspan="3">Другие сферы применения машинного обучения</td><td colspan="2">Некоторые другие сферы применения машинного обучения</td><td colspan="2" rowspan="2">Лишь немногие другие сферы применения машинного обучения</td></tr>
<tr><td colspan="3">РМ3 Интеграция (В5 и В6)</td><td colspan="2">РМ3 Интеграция (В5)</td></tr>
<tr><td colspan="7">Направление работы 1: Поддержка текущих исследований в направлении производства; поддержка новых исследований в рамках других процессов (например, увязки данных) и/или новых источников данных (например, спутниковых данных)</td></tr>
<tr><td rowspan="3">При поддержке</td><td>Качество (точность, своевременность, эффективность, объяснимость и воспроизводимость)</td><td>Качественные данные для обучения</td><td>Навыки/Компетенции</td><td>Вычислительная инфраструктура</td><td>Функциональная совместимость/ Бизнес-процесс</td><td>Этические и правовые нормы</td><td>Безопасность</td></tr>
<tr><td>РМ2 Качество</td><td colspan="3">РМ3 Интеграция (В3 и В4)</td><td colspan="3"></td></tr>
<tr><td>Направление работы 2: Экспериментальное применение практических подходов и методов по некоторым аспектам СКДСА (РМ2);

Направление работы 3: Обзор и совершенствование системы качества</td><td>Направление работы 4: как получить качественные данные для обучения, как поддерживать их в актуальном состоянии, когда переобучать модель, что означает «качественные», как это измерить?</td><td>Направление работы 5: Какие навыки необходимы? Как учиться? Где их найти?</td><td>Подлежит определению</td><td>Подлежит определению</td><td>Направление работы 6: Руководство по этическим вопросам, правила и т. д.</td><td>Подлежит определению</td></tr>
<tr><td rowspan="5">При содействии</td><td colspan="4">Организация</td><td colspan="3">Обмен информацией и сотрудничество</td></tr>
<tr><td colspan="4">РМ3 Интеграция (В1 и В2)</td><td colspan="3">Проект в области машинного обучения ГВУ-МОС</td></tr>
<tr><td colspan="4">Инициативы по ускорению интеграции решений машинного обучения</td><td colspan="3">Исследования и коды МО</td></tr>
<tr><td colspan="4" rowspan="2">Направление работы 7: Создание/поддержание сети руководителей подразделений по интеллектуальной обработке данных;

Направление работы 8: Период после 2021 года: как лучше подготовиться к следующим 2–5 годам? Какие технологии и источники данных мы можем ожидать? Какие навыки нам понадобятся?</td><td colspan="3">Обучение и подготовка</td></tr>
<tr><td colspan="3">Вебинар по проекту ГВУ-МОС</td></tr>
</table>