ИСПОЛЬЗОВАНИЕ ВОЗМОЖНОСТЕЙ ИИ И МАШИННОГО ОБУЧЕНИЯ В ОФИЦИАЛЬНОЙ СТАТИСТИКЕ: ОПЫТ SORS

Трансформация работы со статистическими данными посредством передовых технологий на базе Статистического управления Республики Сербия (SORS)

Марко Груичич
Адиль Колакович

Вводная информация о применении ИИ и машинного обучения в официальной статистике

- ИИ и машинное обучение совершают настоящую революцию в официальной статистике, автоматизируя сложные процессы, повышая точность данных и обеспечивая анализ в режиме реального времени.

- Эти технологии прошли путь от базовых моделей машинного обучения до сложных систем, таких как большие языковые модели (LLM).

- Интеграция ИИ/машинного обучения не только повышает эффективность, но и открывает новые возможности для использования нетрадиционных источников данных, способствуя своевременному получению статистических данных.

Обзор цифровой переписи SORS

- Перепись 2022 года стала первой полностью цифровой переписью в Сербии, в которой ИИ и машинное обучение интегрировались на всех этапах сбора и обработки данных.

- SORS удалось улучшить точность данных, снизить затраты и ускорить процесс переписи за счет применения цифровых технологий.

- Этот переход на цифровые технологии установил новый стандарт для проведения переписей в Сербии, продемонстрировав потенциал ИИ и машинного обучения, которые в корне изменили статистическую практику.

Машинное обучение на этапе обработки материалов переписи

- SORS использовало машинное обучение для классификации профессий и экономической деятельности на основе классификаций ISCO и NACE.

- Машинное обучение автоматизировало традиционно трудоемкий процесс классификации, обеспечив более быстрые и точные результаты.

- Модели были обучены на больших наборах данных, которые включали такие переменные, как образование, возраст и пол, что повышало точность классификации.

Обезличивание данных и обработка в облаке

- SORS использовало облачную инфраструктуру для эффективной обработки больших наборов данных для реализации задач машинного обучения, обеспечивая масштабируемость и защиту данных.

- Перед обработкой в облаке данные обезличивались для защиты конфиденциальной информации и охраны частной жизни респондентов.

- Облачная среда, оснащенная высокопроизводительным оборудованием, позволила SORS быстро обрабатывать сложные алгоритмы машинного обучения.

Обучение моделей МО и источники данных

- Обучение началось с данных переписи 2011 года, и было дополнено данными текущих обследований, такими как обследование рабочей силы, и данными Центрального регистр обязательного социального страхования (административный источник в Сербии).

- Эти разнообразные источники данных сформировали надежную обучающую базу для точной классификации занятий и видов деятельности.

- Объединение нескольких источников данных позволило SORS создать надежную модель машинного обучения, устраняющую потенциальные ошибки, связанные с различными демографическими факторами.

Роль IST в повышении эффективности ИИ/машинного обучения

- Эффективно интегрированный сбор данных:

Расширенные функции IST, включая эффективный сбор данных и надежный логический контроль, обеспечили доступность высококачественных и «чистых» данных для обработки с применением технологий ИИ/машинного обучения.

- Современная архитектура:

Архитектура IST, построенная на базе сервера MS SQL DB Server, обеспечивает стабильную и масштабируемую основу, сводя к минимуму необходимость обширной предварительной обработки перед применением технологий ИИ/машинного обучения.

- Улучшенная эффективность ИИ/машинного обучения:

Начав с хорошо структурированных и точных данных, система IST значительно повысила производительность ИИ/машинного обучения, сократив требования к ресурсам и обеспечив осуществимость проекта.

- Исключение принципа GIGO:

Возможности IST помогли нам избежать распространенной ошибки «Каков вопрос, таков ответ», что привело к получению более достоверной и практически полезной аналитической информации на основе наших моделей ИИ/машинного обучения

Использованные алгоритмы машинного обучения

- SORS протестировало и выбрало классификатор Random Forest («метод случайного леса») из-за его превосходной точности при классификации занятий и видов деятельности.

- Модель «Метод случайного леса» обеспечила точность на уровне 98 %, что является решающим фактором в получении высококачественных данных переписи.

- Данный алгоритм идеально подходил для обработки разнообразных и сложных данных, собранных в ходе переписи.

Достижения в области точности классификации

- Классификатор Random Forest повысил точность классификации занятий и видов деятельности, обеспечив уровень точности 98 %.

- Эта высокая точность была подтверждена на образцах, закодированных вручную, что обеспечило надежные результаты.

- Описанные улучшения повысили общее качество данных переписи, обеспечив надежную основу для будущих статистических операций.

Практические преимущества применения ИИ и машинного обучения при проведении переписи

- ИИ и машинное обучение сократили время обработки и количество ошибок за счет автоматизации задач классификации, обеспечивая стабильное качество данных.

- Эти технологии позволили SORS более результативно обрабатывать большие наборы данных, делая процесс переписи более быстрым и затратоэффективным.

- Интеграция ИИ и машинного обучения в перепись подчеркнула их потенциал для более широкого применения в официальной статистике.

Трудности и полученный практический опыт

- Трудности включали проблемы с качеством данных в текстах после применения технологии OCR и были связаны с требованиями к вычислительной мощности для обучения модели МО.

- SORS разработало вспомогательные модели машинного обучения для исправления текста и оптимизировала облачные ресурсы для решения подобных проблем.

- Эти решения послужили ценными уроками для будущих проектов, снабдив SORS знаниями для расширения использования возможностей ИИ и машинного обучения.

Применение технологий ИИ и машинного обучения в национальных статических управления (NSO) на глобальном уровне: Статистическое управление Канады

- Статистическое управление Канады использует машинное обучение для автоматизации кодирования ответов на опросы, что приводит к сокращению времени обработки и повышению точности данных.

- Машинное обучение сократило необходимость ручного кодирования, минимизировало ошибки и оптимизировало операции.

- Данный подход демонстрирует универсальность и эффективность машинного обучения в статистической сфере.

Применение технологий ИИ и машинного обучения в национальных статических управления (NSO) на глобальном уровне: Управление национальной статистики (Великобритания)

- Управление национальной статистики (ONS) в Великобритании использовало ИИ для классификации видов коммерческой деятельности, тем самым повысив точность и своевременность экономической статистики.

- Применение технологий ИИ позволило ONS более эффективно обрабатывать и анализировать большие наборы данных, предоставляя надежные данные для корректировки экономической политики.

- Этот успех подчеркивает потенциал ИИ, способный произвести революцию в сборе и обработке статистических данных.

Дальнейшие направления развития ИИ и машинного обучения, как их видит SORS

- SORS планирует распространить использование ИИ/машинного обучения на другие области статистики, включая экологическую и социальную статистику.

- Будущие инициативы будут включать сотрудничество с другими NSO и международными организациями для обмена передовым опытом.

- SORS стремится внедрять инновации в области ИИ и машинного обучения в статистике, повышая точность, эффективность и актуальность данных.

Этические аспекты и передовой опыт

- Соблюдение глобальных этических стандартов обеспечивает ответственное и прозрачное использование ИИ/машинного обучения в официальной статистике.

- SORS разработало руководящие принципы этического развертывания ИИ, уделяя особое внимание справедливости, подотчетности и прозрачности.

- Эти руководящие принципы предотвращают предвзятость, охраняют частную жизнь и позволяет всем заинтересованным сторонам статистического процесса в полной мере использовать преимущества ИИ.

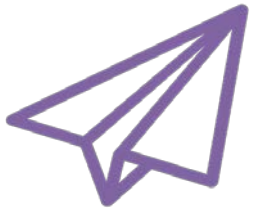
Заключение. Вопросы и ответы

- Перепись SORS 2022 года продемонстрировала преобразующую силу ИИ/машинного обучения в официальной статистике, установив новый стандарт для сбора цифровых данных.

- SORS стремится продолжать внедрять инновации в области ИИ/машинного обучения, работая с глобальными партнерами над продвижением этих технологий в статистике.

- В данной презентации были освещены достижения и трудности, с которыми пришлось столкнуться SORS в области ИИ/машинного обучения, и мы приветствуем любые вопросы и возможности для обсуждения.

Контактная информация



Эл. почта: ist@stat.gov.rs

Веб-сайт: https://istportal.net/