

Практическое задание №1

Цель задания:

В этом задании Вы решите простую и во многом учебную задачу по созданию автоматического конвейера проекта машинного обучения. Подобный подход с применением простых скриптов автоматизации для “склейки” отдельных частей конвейера используется в небольших проектах.

Содержание задания:

Необходимо из “подручных средств” создать простейший конвейер для автоматизации работы с моделью машинного обучения на Linux. Отдельные этапы конвейера машинного обучения описываются в разных python-скриптах, которые потом соединяются (иногда используют термин “склеиваются”) с помощью bash-скрипта.

Этапы:

1. Создайте python-скрипт (data_creation.py), который создает различные наборы данных, описывающие некий процесс (например, изменение дневной температуры). Таких наборов должно быть несколько, в некоторые данные можно включить аномалии или шумы. Часть наборов данных должны быть сохранены в папке “train”, другая часть в папке “test”.
2. создайте python-скрипт (model_preprocessing.py), который выполняет предобработку данных, например, с помощью `sklearn.preprocessing.StandardScaler`.
3. создайте python-скрипт (model_preparation.py), который создает и обучает модель машинного обучения на построенных данных из папки “train”.
4. создайте python-скрипт (model_testing.py), проверяющий модель машинного обучения на построенных данных из папки “test”.
5. Напишите bash-скрипт (pipeline.sh), последовательно запускающий все python-скрипты.

Подготовленные скрипты необходимо опубликовать в git репозитории, ссылку на который необходимо предоставить как результат выполнения задания. На портал загрузите txt-файл, названный по ФИО, с такой ссылкой.

Творческое задание (выполняется по желанию)

Предположим, что заказчик поставил перед командой задачи: разработать прогнозную модель для зарплат специалистов MLOps на два года вперед. Заказчик хочет Определять уровень зарплат в среднем по рынку, чтобы нанимать в течении двух лет 10 специалистов и не переплатить более 500.000 т.р. (суммарно). Решение заказчик хочет эксплуатировать на своих серверах (поскольку не хочет раскрывать персональные данных своих сотрудников). Однако сервера у заказчика ненадежные, технические описания он не предоставляет.

1. Предложите технические метрики машинного обучения, которые могли бы соответствовать этой постановке задачи и инструменты для ее решения.
2. С какими сложностями может столкнуться специалист MLOps при реализации задач такой постановке?
3. Какой перечень мероприятий можно предложить для минимизации рисков?