

Конкурсное задание

REASkills 2025

Машинное обучение и большие данные

С3

Модуль 5 Разработка решения для задачи с

большими данными. Разработка модели

СОДЕРЖАНИЕ

Модуль 5 данного Конкурсного задания состоит из следующей документации / файлов:

1. C3_M5.docx (Инструкция к пятому модулю)

ВВЕДЕНИЕ

В этом модуле Вам предстоит разработать модель машинного обучения, способную хорошо работать на больших данных.

ИНСТРУКЦИЯ УЧАСТНИКУ

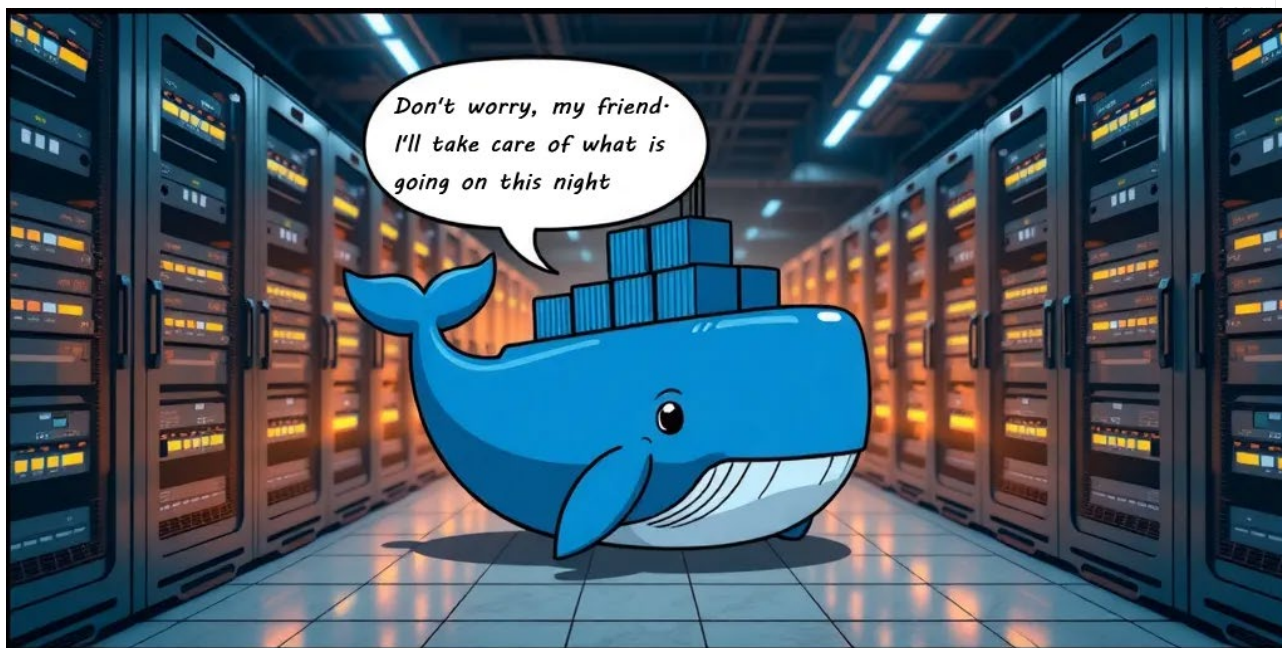


К концу этого модуля у вас должны быть достигнуты следующие результаты:

ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

1 Разработка модели машинного обучения

1. Обучающая выборка данных разделена на обучающую и тестовую выборку
2. Испытано как минимум три алгоритма разной сущности (RandomForest и XGBoost – разные сущности, а XGBoost и GradientBoosting – нет). Для оценки точности следует использовать **тестовую** выборку
3. Испытанные алгоритмы сравнены по точности, а также по скорости работы
4. Лучший алгоритм реализован в виде пайплайна, принимающего на вход пути к файлам, а на выходе возвращает обученную модель, а также метрики модели на **валидационной** выборке в виде словаря (необходимые метрики: средний коэффициент детерминации R2 и средний RMSE).



2 Контейнеризация обучения

1. Установлен доступ на удаленный сервер через ssh:
 - 1) прописать команду для переноса пароля (чтобы не надо было его постоянно вводить)


```
ssh-copy-id c[X]@10.0.10.[Y]
```

 где [X] – номер участника
[Y] – номер виртуальной машины (на бумажке)
 - 2) прописать команду для подключения к серверу:


```
ssh c[X]@10.0.10.[Y]
```

 где [X] – номер участника
[Y] – номер виртуальной машины (на бумажке)
2. Выберите способ создания выборки данных на удаленном сервере исходя из цели минимизации временных затрат.
 - ЛИБО На сервер в директорию /home/user/bigdata/data скопированы файлы data_X, data_y, data_X_val, data_y_val при помощи утилиты scp или аналога. Не забудьте сначала создать требуемые директории
 - ЛИБО подготовлен и запущен скрипт, который проведет предобработку данных (на сервере у вас уже будут подготовлены файлы df_{x} и target_{x}, а также df_val и target_val) на сервере и создаст файлы data_X, data_y, data_X_val, data_y_val как результат такой обработки
3. Подготовлен .py файл для контейнеризации. Файл должен содержать все необходимое для запуска пайплайна из предыдущего раздела. После обучения скрипт должен сохранить обученную модель в рабочую директорию контейнера в сериализованном виде (pickle-формат), с названием файла model.pkl. Также скрипт сохраняет метрики, полученные на валидационной выборке и временную метку запуска в текстовый файл (мини-отчет) metrics.txt в рабочую директорию контейнера.
4. Подготовлен Dockerfile для создания контейнера, предназначенного для обучения модели на удаленном сервере. В Dockerfile следует указать, какую версию питона следует установить (или принять за базовый образ), а также какой минимально необходимый набор библиотек установить.
Также следует прописать базовую директорию, команду для копирования скрипта .py в рабочую директорию контейнера и указания для запуска скрипта с обучающим пайплайном.
5. Подготовлен .dockerignore файл, куда прописана директория ./BIG_DATA (/home/c[X]/BIG_DATA)
6. Файлы Dockerfile, .py файл и .dockerignore переданы на удаленный сервер в директорию /home/c[X]
7. На удаленном сервере запущена сборка контейнера в образ
8. На удаленном сервере запущен контейнер с монтированием к нему директории /home/c[X]/BIG_DATA в рабочую папку контейнера

В случае использования контейнеризации вам разрешается использовать вечернее и ночное время для вычислений на удаленном сервере (не на своей локальной машине!)

Примечание 1: Текст, выделенный курсивом является рекомендательным порядком действий. В случае использования контейнеризации участники могут заработать баллы только за качество (точность) модели, а также за отчет и мини-отчет. Способы достижения результата могут быть разные

Примечание 2: В случае, если вы не сможете победить контейнеризацию – обучите модель локально в рамках времени модуля и отразите результаты (метрики) в отчете. Результаты данного модуля существенно зависят именно от точности модели

3 ОТЧЕТ

1. Предоставлен отчет о проделанной работе. Внимание: оценка сессии будет проводиться на основании отчета. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода и результат работы программы.
2. Отчёт должен быть предоставлен в папке на рабочем столе /home/c[X]/Рабочий\ стол/C[X]_M5, где [X] – номер рабочего места. Папка должна содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. Обязательно наличие файлов:
 - Jupyter Notebook C[X]_M5.ipynb (или аналог – с возможностью запустить и исполнить участки кода),
 - C[X]_M5.HTML (или PDF), где [X] – номер команды (участника).
 - Python файл в случае обучения модели на удаленном сервере
 - Dockerfile
 - .dockerignore файл
 - Metrics.txt мини-отчет