**Anna Guseva wrote:**

Короткое видео по пайплайну:

https://www.coursera.org/lecture/ml-pipelines-bert/creating-machine-learning-pipelines-S5u1v

и вот небольшая статья:

https://medium.com/analytics-vidhya/what-is-a-pipeline-in-machine-learning-how-to-create-one-bda91d0ceaca

просто сейчас статьи, которые я нахожу, они какие-то не обобщенные, а с деталями под конкретную задачу

но думаю можно начать отсюда

Graphical user interface

Description automatically generated

вот это очень обобщенная картинка пайплайна - в целом, все что вам нужно знать без каких-либо ДС (дата сайенс) деталей

**Anna Guseva, [23 Mar 2022, 11:55:12]:**

т.е., как я говорила, первая часть это идет разработка самой модели

- мне на вход приходят сырые данные, а потом, в зависимости от области задачи, я анализирую данные, выбираю какие трансформации с ними сделать, какие данные оставить, а какие удалить

- на следущем этапе выбираю какие модели могу использовать, т.е. нахожу какие модели и с какими параметрами дадут наилучших предикт

- тренирую выбранные модели

- оцениваю точность

после этого этапа что я получаю для деплоя:

- код для предобработки данных, чтобы можно было подавать в модель

- архитектура выбранной/ых моделей

- веса/параметры моделей

как это используется в продакшене (часть, которую мы используем в челендже):

- считываю данные

- подготавливаю данные для чтения модели

- считываю/вызываю натренированную модель со всеми параметрами

- подаю данные в модель и получаю от нее ответ в виде предикта

думаю, будут более понятно в простом примере который подготовлю

**и тут вопрос про тестировку:**

как по мне тестировать можно две разных вещи

- саму реализация, т.е. что будет, если подавать пустоту, какие-то граничные значения, данные другого формата, большое кол-во и прочее

- правильность результатов модели (т.е. у нас точность в большинстве случаев не 100% и нужно отлавливать какие именно случаи плохо отрабатываются). Например, задачи классификации объекта на изображении - можно найти что объект плохо классиф, если оно перекрыто другим объектом и прочее

но, я думаю, что в этом челендже для тестировки второй вариант не будем рассматривать, т.к. нам точность модели не важна

но это то, что я из головы взяла, думаю можно еще разных туториалов по тестировки посмотреть

**Разварачивания Flask для DataScience проектов**

Два небольших туториала, может поможет:

* <https://towardsdatascience.com/how-to-build-a-machine-learning-api-using-flask-2fb345518801>
* <https://www.statworx.com/en/content-hub/blog/how-to-build-a-machine-learning-api-with-python-and-flask/>