Открыт приём заявок!

Первое ДЗ

Машинное обучение в продакшене

Всем привет!

Выполнение первого домашнего задания оценивается в 30 баллов.

Вам потребуется сделать "production ready" проект для решения задачи классификации, то есть написать код для обучения и предикта, покрыть его тестами, обучить с помощью него модельку.

Для обучения модели, можете использовать датасет https://www.kaggle.com/ronitf/heart-disease-uci, либо иной небольшой датасет для классификации(если используете другой, то опишите какой и как его получить в readme)

Пример подобного разбирали на паре https://github.com/made-ml-in-prod-2021/ml_project_example, не нужно отсюда копировать=)

Критерии (указаны максимальные баллы, по каждому критерию ревьюер может поставить баллы частично):
-2) Назовите ветку homework1 (1 балл)
-1) положите код в папку ml_project
0) В описании к пулл реквесту описаны основные "архитектурные" и тактические решения, которые сделаны в вашей работе. В общем, описание что именно вы сделали и для чего, чтобы вашим ревьюерам было легче понять ваш код. (3 балла)

1) Выполнение EDA, закоммитьте ноутбук в папку с ноутбуками (3 баллов)
Вы так же можете построить в ноутбуке прототип(если это вписывается в ваш стиль работы)
Можете использовать не ноутбук, а скрипт, который сгенерит отчет, закоммитьте и скрипт и отчет (за это + 1 балл)

2) Проект имеет модульную структуру(не все в одном файле =)) (3 баллов)

3) использованы логгеры (2 балла)

4) написаны тесты на отдельные модули и на прогон всего пайплайна(5 баллов)

5) Для тестов генерируются синтетические данные, приближенные к реальным (5 баллов)
- можно посмотреть на библиотеки https://faker.readthedocs.io/en/, https://feature-forge.readthedocs.io/en/latest/
- можно просто руками посоздавать данных, собственноручно написанными функциями
как альтернатива, можно закоммитить файл с подмножеством трейна(это не оценивается)

6) Обучение модели конфигурируется с помощью конфигов в json или yaml, закоммитьте как минимум 2 корректные конфигурации, с помощью которых можно обучить модель (разные модели, стратегии split, preprocessing) (2 балла)

7) Используются датаклассы для сущностей из конфига, а не голые dict (3 балла)

8) Используйте кастомный трансформер(написанный своими руками) и протестируйте его(3 балла)

9) Обучите модель, запишите в readme как это предлагается (3 балла)

10) напишите функцию predict, которая примет на вход артефакт/ы от обучения, тестовую выборку(без меток) и запишет предикт, напишите в readme как это сделать (3 балла)

11) Используется hydra (https://hydra.cc/docs/intro/) (3 балла - доп баллы)

12) Настроен CI(прогон тестов, линтера) на основе github actions (3 балла - доп баллы (будем проходить дальше в курсе, но если есть желание поразбираться - welcome)
13) Проведите самооценку, опишите, в какое колво баллов по вашему мнению стоит оценить вашу работу и почему (1 балл доп баллы)

PS: Можно использовать cookiecutter-data-science https://drivendata.github.io/cookiecutter-data-science/ , но поудаляйте папки, в которые вы не вносили изменения, чтобы не затруднять ревью

Сроки выполнения:
Мягкий дедлайн: 2 мая 23:59
Жесткий дедлайн: 9 мая 23:59
(небольшие доработки по требованиям ревьюеров можно будет вносить после срока, важно, чтобы основная часть кода была написана до дедлайна)
После мягкого дедлайна все полученные баллы умножаются на 0.5

Процедура сдачи:
После выполнения ДЗ, создаем пулл реквест, в ревьюеры добавляем teachers и students(см. скрины), к вам в ревьюеры автоматически добавиться один из "профессиональных проверяющий" и случайный студент. Оценка преподавателя будет использована как финальная. Преподаватель начнет проверять работу через 2-3 дня после добавления, чтобы дать возможность высказаться проверяющему студенту.

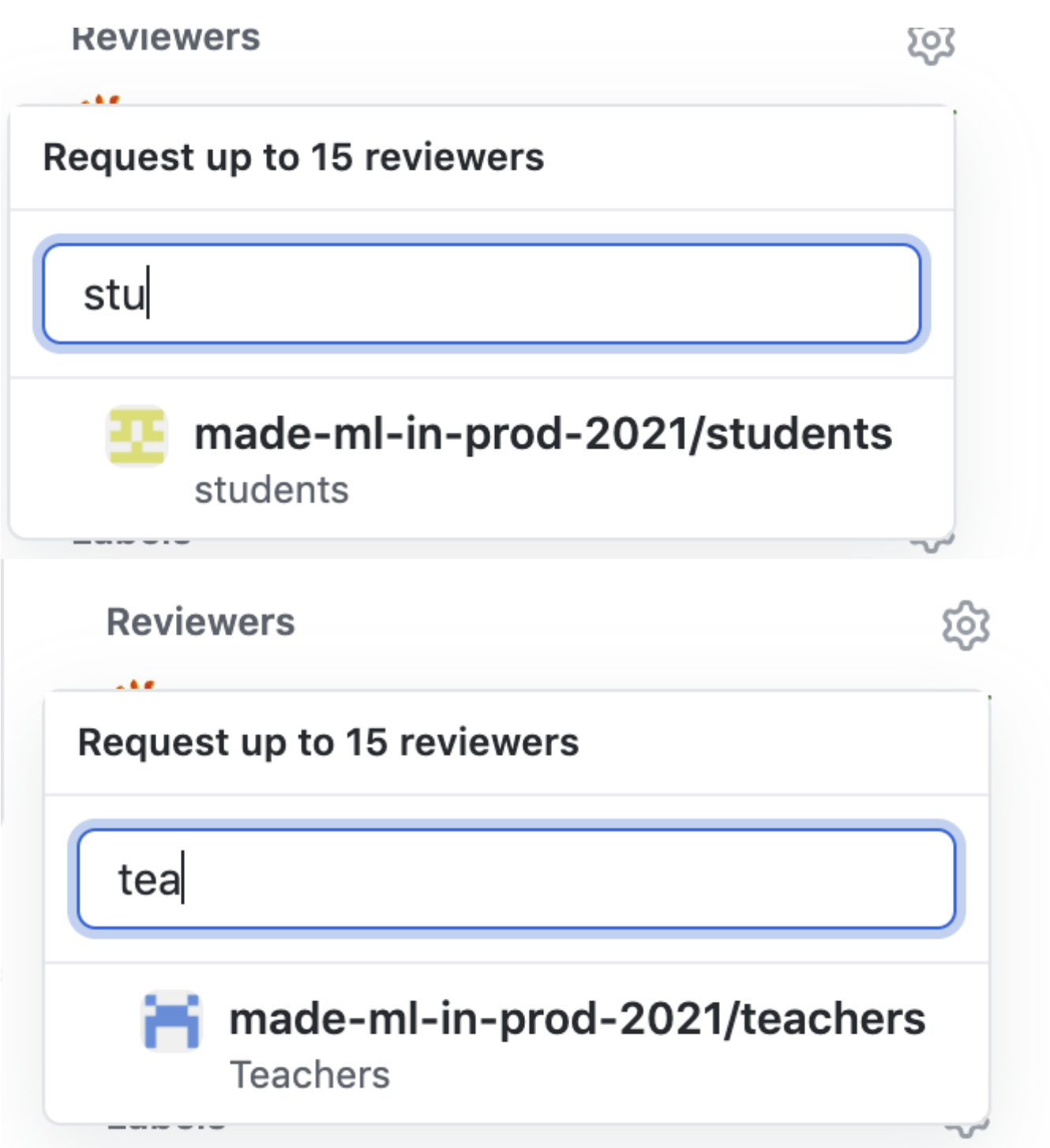
Про ревью друг друга:
Когда вас добавили в ревью, посмотрите код коллеги, напишите свои замечания, если что-то по вашему мнению сделано не очень хорошо, что можно в этом коде улучшить, если видите особенно клевый участок кода, то можно похвалить автора=) И не стоит себя недооценивать! Далее напишите, какому кол-ву баллов по вашему мнению соотвествует работа и почему бы вы лично что-то сняли.

Если вас назначили и вы **НЕ БУДЕТЕ** делать ревью, смотреть код принципиально, то удалите себя из ревьюеров и добавьте students заново(выберет другого)

Если решили покинуть курс(не сдавать домашние работы), то заполните форму, я удалю из организации и вы не будете попадать в ревьюеры. (не вводите ники друг друга, это жестко!)

Какая мотивация для вас так делать?
1) Фан
2) 2 балла за факт сделанного ревью и поставленной оценки
3) Каждый из преподавателей для каждого домашнего задания будут выбирать лучшего ревьюера и мы поставим ему не 2, а 7 баллов(не обязательно количество таких людей будет равно количеству преподавателей, но так как вас много -- это вполне возможно)

Всем хороших выходных!



Норм?

слишком изи

норм

сложна

Воздержаться