Финальный проект

Где брать данные

а) Свой проект, в котором можно применить полученные знания (например, с работы)
б) Любое соревнование из https://www.kaggle.com/competitions, в котором есть призовой фонд. Можно уже завершенные. Они далеко не все похожи на то, что мы проходим, так что придется поискать интересное. В каких-то машинное зрение, в каких-то текстовые задачи. Если возьмете то, что не проходили, это не будет минусом. Несколько примеров:

- https://www.kaggle.com/c/ieee-fraud-detection
- https://www.kaggle.com/c/tmdb-box-office-prediction
- https://www.kaggle.com/c/petfinder-adoption-prediction
- https://www.kaggle.com/c/m5-forecasting-accuracy

Копирование кода

Нормальной практикой считается копировать код из чужих ноутбуков, интернета, наших семинаров и т.д.
Единственное, если Вы берете соревнование с kaggle, то не стоит полностью копировать чужое решение с этого же соревнования. Скомпилировать несколько чужих ноутбуков в один - это ок.

Главное - понимать, что делает Ваш код. Я буду задавать вопросы по коду на понимание: больше всего вопросов для простых и популярных соревнований с kaggle, и меньше всего для своих проектов.

Что должно быть в проекте

Ноутбук, который мы разбирали https://www.kaggle.com/dejavu23/titanic-survival-seaborn-and-ensembles можно использовать как шаблон. Только, так много, как там, делать не обязательно.

Требуемый объем. 3-5 графиков, 2-3 модели, и выводы после каждого эксперимента.

Ниже общий план того, что должно быть в проекте

1. Постановка задачи

   Ответить на вопросы (можно устно):
   - Какая целевая переменная (что предсказываем)
   - Задача классификации или регрессии (или кластеризации, или другая)?

- Какую метрику будем использовать для оценки качества (на kaggle это есть в разделе соревнование->overview->evaluation)
- 2 балла

2. Анализ данных

- проанализировать целевую переменную: как она распределена, сбалансированы ли классы (если классификация)
- сколько у нас данных и какие есть признаки
- проанализировать основные зависимости между признаками и целевой переменной, попытаться понять, какие признаки будут значимыми для модели, а какие - нет (это не значит, что их сразу нужно выкинуть)
- 4 балла

3. Подготовка признаков

- избавляемся от пропусков, если есть
- добавляем новые значимые признаки, если нужно (feature engineering)
- превращаем признаки в числа (для категориальных признаков будет лучше использовать one-hot encoding)
- любые дополнительные действия, которые посчитаете нужными (scaling, понижение размерности)
- 2 балла

4. Обучение модели и валидация

- попробовать обучить несколько моделей. Обычно, или случайный лес работает намного лучше, или линейная модель.
- измерить качество моделей, сделать выводы
    - обратите внимание, нужно получить реалистичную оценку качества (несмещенную). Например, с помощью кросс-валидации
- любые дальнейшие эксперименты на Ваше усмотрение
- 2 балла

Формат сдачи и оценка

В результате у Вас получится ноутбук с кодом и графиками.
При сдаче Вы будете показывать мне графики и код и рассказывать, что происходит. В процессе буду задавать вопросы. Если Вам интересно будет и дальше работать с проектом, то и Вы мне задавайте вопросы по своему проекту :)

Максимальная оценка - 10
За что снижается оценка:

- Если слушатель не понимает, что делает код, который он скопировал
- Если отсутствует одна из важных частей плана (например, не дошли до обучения модели)
- Если допущена совсем очевидная ошибка. Например, не удалено имя пользователя из признаков (при этом, слушатель понимает, что оно там быть не должно). Или целевая переменная присутствует среди признаков, и модель предсказывает то, что ей и так дано (да, так уже было:)).

За что задаются доп. вопросы

- Если очень мало сделано
- Если много кода скопировано из того-же соревнования на kaggle, которое Вы взяли

Доп. вопросы - несложные вопросы по всему курсу, которые позволят получить оценку выше

Оценка *не снижается*, если получено низкое качество, и причина не в том, что вы что-то сделали не так, а просто в данных. Отрицательный результат - тоже результат. Или, если задача реально сложная и в рамках учебного проекта пока не получилось достичь высокого качества.

Примеры проектов на kaggle

Если Вы сделали проект на kaggle, то можете опубликовать его (сделать видимым для всех). Это может принести Вам upvote'ы, если ноутбук окажется актуальным: например, новое соревнование, где Вы сделали один из первых подобных ноутбуков.

Вот некоторые опубликованные работы из прошлых групп:

- https://www.kaggle.com/polmast/demin-av-eee-cis
- https://www.kaggle.com/pavlovivan/ds-75
- https://www.kaggle.com/loider/kernel51fe621948
- https://www.kaggle.com/olegkokhanskiy/crossfit-games-2019-starting-analysis