Участие в соревнованиях

Основные шаги работы над задачей

1. Визуализация данных
2. Очистка данных и генерация признаков
3. Разработка схемы локального контроля
4. Построение предсказательной модели

Визуализация данных

Зачем? Оценить масштабы задачи, посмотреть на вариативность в данных, проанализировать признаки и объекты, оценить необходимость предобработки данных, найти новые признаки, подумать о применимости моделей...
Инструменты: matplotlib, pandas, seaborn, etc.
Ресурсы:
Презентация
Обзор инструментов и статей
Скрипты Kaggle

Очистка данных и генерация признаков

Зачем? Получить матрицу объекты-признаки, с которой смогут работать метода sklearn; улучшить качество решения задачи.
Инструменты: sklearn.preprocessing, pandas
Ресурсы:
Статья на Хабрахабре
Будет лекция про признаки

Разработка схемы локального контроля

Зачем? Подбирать гиперпараметры моделей; отбирать модели, признаки, объекты (!)...
Инструменты: sklearn.cross_validation, numpy, pandas
Особенности:

- выбор объема отложенной выборки и числа блоков в кросс-валидации
- учитывать временную ось в данных, если она есть
- одинаковое распределение данных в обучении и в контроле

Построение предсказательной модели

- отбор и настройка моделей
- композиции

Public и Private LB

Не переобучиться!

- тестировать модели на локальном контроле
- не выбирать модель по public LB, особенно если данных не очень много