Project Check List

1. Вы выбрали данные, в них как минимум 1 целевая переменная и 9 признаков, которые хотя бы теоретически с ней связаны.

2. Вы определили тип целевой переменной и тип задачи, которую решаете (если непрерывная - задача регрессии, если категориальная - классификации).

3. Разбиватае ваши данные на train и test.

4. Вы построили минимум пять визуализации (из них три разного типа), которые отражают наличие или отсуствия связи ваших признаков с зависимой переменной.

5. Преобразовали одну количественную переменную в категориальный признак или уменьшили размерность категориальной переменной.

6. На основе EDA сформулировали гипотезы о связи переменных. На основе EDA отобрали признаки, которые будете использовать в моделях.

7. Выбираете метрику измерения ошибки. Если у вас задача классификации, то accuracy, если регрессии - то либо mean absolute error либо mean squared error. Это из тех, что мы проходили и которые вы должны знать. Если знаете другие (а они есть и они лучше), можете взять их. MAE берите, если в вашей целевой переменной есть выбросы.

8. На основе вашей метрики строите baseline модель. Для задачи классификации - проверяете ошибку для тестовой выборки, если бы вы все заполнили самым часто встречающимся классом. Для задачи регрессии - предсказываете все значения медианой, если выбрали MAE, или средним, если выбрали MSE. Считаете метрику ошибки.

9. Выбираете три модели - регрессоры или классификаторы, в зависимости от задачи. Оцениваете качество предсказания модели без настроек. Будьте готовы в общих чертах объяснить, как работает этот алгоритм. Тут прямо в общих - общий принцип работы решающего дерева, как работает kNN. Завтра на занятии можем обсудить.

10. Для каждой модели реализуете поиск параметров по сетке (на кросс-валидации (https://github.com/rogovich/2019-2020_PolSci_Data_Analysis_in_Python/blob/master/14week_LDA_Titanic/14week_Model_Selection_Full.ipynb) для минимум двух параметров. Будьте готовы ответить на вопрос, за что отвечает этот параметр. Не выбирайте параметры, которые совсем не понимаете.

11. Выберите лучшие параметры для каждой из трех моделей и сравните коэффицент ошибке на отложенной выборке (test). Улучшилось ли предсказание по сравнению с baseline и насколько? Подтвердились ли гипотезы (хорошие ли получились модели)?