Курсовой проект.

Ссылка на данные https://www.kaggle.com/c/realestatepriceprediction

Сдача проекта:

1.Прислать в раздел Задания Урока 10 ("Вебинар. Консультация по итоговому проекту") ссылку на программу в github (программа должна содержаться в файле Jupyter Notebook с расширением ipynb). (Pull request не нужен, только ссылка ведущая на сам скрипт).

2.Приложить файл с названием по образцу NVBaranov\_predictions.csv с предсказанными ценами для квартир из test.csv (файл должен содержать два поля: Id, Price). В файле с предсказаниями должна быть 5001 строка (названия колонок + 5000 предсказаний).

Сроки и условия сдачи: Дедлайн: сдать проект нужно в течение 72 часов после завершения Урока 10 ("Вебинар. Консультация по итоговому проекту"). Для успешной сдачи должны быть все предсказания (для 5000 квартир) и R2 должен быть больше 0.6. При сдаче до дедлайна результат проекта может попасть в топ лучших результатов. Повторная сдача и проверка результатов возможны только при условии предыдущей неуспешной сдачи. Успешный проект нельзя пересдать в целях повышения результата. Проекты, сданные после дедлайна или сданные повторно, не попадают в топ лучших результатов, но можно узнать результат. В качестве итогового результата берется первый успешный результат, последующие успешные результаты не учитываются.

Примечание: Все файлы csv должны содержать названия полей (header - то есть "шапку"), разделитель - запятая. В файлах не должны содержаться индексы из датафрейма.

Рекомендации для файла с кодом (ipynb):

1.Файл должен содержать заголовки и комментарии

2.Повторяющиеся операции лучше оформлять в виде функций

3.Не делать вывод большого количества строк таблиц (5-10 достаточно)

4.По возможности добавлять графики, описывающие данные (около 3-5)

5.Добавлять только лучшую модель, то есть не включать в код все варианты решения проекта

6.Скрипт проекта должен отрабатывать от начала и до конца (от загрузки данных до выгрузки предсказаний)

7.Весь проект должен быть в одном скрипте (файл ipynb).

8.При использовании статистик (среднее, медиана и т.д.) в качестве признаков, лучше считать их на трейне, и потом на валидационных и тестовых данных не считать статистики заново, а брать их с трейна. Если хватает знаний, можно использовать кросс-валидацию, но для сдачи этого проекта достаточно разбить данные из train.csv на train и valid.

9.Проект должен полностью отрабатывать за разумное время (не больше 10 минут), поэтому в финальный вариант лучше не включать GridSearch с перебором большого количества сочетаний параметров.

10.Допускается применение любых моделей машинного обучения из библиотеки sklearn.

11.Решение не должно полностью повторять этот скрипт

Прогнозирование на тестовом датасете

1.Выполнить для тестового датасета те же этапы обработки и постронияния признаков (лучше выполнять действия сразу для двух датасетов)

2.Не потерять и не перемешать индексы от примеров при построении прогнозов

3.Прогнозы должны быть для все примеров из тестового датасета (для всех строк)