Подготовка данных и ML. Домашнее задание

Задача состоит в подготовке данных к использованию в моделях машинного обучения и прогнозировании целевого параметра.

Описание исходных данных

На выбор вам дается 2 датасета:

- Данные о стоимости мобильных телефонов – homework3_smartphones.csv
- Данные о стоимости автомобилей – homework3_cars.csv

Соответственно, в первом случае необходимо спрогнозировать стоимость мобильных телефонов, а во втором – стоимость автомобилей.

Постановка задачи

Постарайтесь по максимуму использовать все стадии подготовки данных. Понятно, что если нет пропусков или выбросов, то ничего чистить не нужно, но по крайней мере нужно все проверить.

Итак, необходимо выполнить следующие шаги:

1. Очистка данных.

   Обработка пропусков, выбросов, дубликатов.

2. Предобработка данных.

   Нормализация или стандартизация данных, кодирование.

3. Трансформация данных.

   Feature engineering. Очень часто среди текстовых признаков скрывается очень ценная информация. Для ее вытаскивания можно использовать стандартные операции со строками (split, strip, lower и т.п.), модуль для работы с регулярными выражениями (https://tproger.ru/translations/regular-expression-python).

   Не забывайте про удаление скоррелированных признаков и признаков, которые явно не будут иметь влияние на целевой признак.

4. Обучение модели и прогнозирование результатов.
   Для сравнения результатов необходимо применить для задачи две разных модели.
   Не забывайте на разбиение выборки на тренировочную и тестовую. Постарайтесь, чтобы распределение целевого признака было примерно равным в обеих выборках.
   Вы можете выбрать любые модели машинного обучения с учителем из Sklearn (https://scikit-learn.org/stable/supervised_learning.html), решающие задачи регрессии.
5. Оценка результатов
   Сравните результаты работы двух моделей и сделайте на основании результатов сравнения вывод.
   Для сравнения используйте разные метрики: MSE, RMSE, MAPE, MAE и любые другие.

Оформление результатов

Результат работы должен быть оформлен в формате Jupyter-ноутбука.

Основные стадии должны быть выделены заголовками, операции над данными, анализ данных и другие действия должны сопровождаться краткими комментариями.