Предсказать price - цену автомобиля на основе его характеристик

Данные:
Данные представляют собой информацию о различных автомобилях, где каждая строка соответствует отдельному автомобилю с различными характеристиками. Описание колонок:

1. car_ID — уникальный идентификатор автомобиля.
2. symboling — индекс безопасности автомобиля (числовое значение, где более высокие значения могут означать более высокий риск).
3. CarName — название модели автомобиля.
4. fueltype — тип топлива, используемого автомобилем (например, "gas" — бензин, "diesel" — дизель).
5. aspiration — тип системы впуска (например, "std" — стандартная система, "turbo" — турбонаддув).
6. doornumber — количество дверей в автомобиле (например, "two" или "four").
7. carbody — тип кузова автомобиля (например, "convertible" — кабриолет, "sedan" — седан).
8. drivewheel — тип привода колес (например, "rwd" — задний привод, "fwd" — передний привод, "4wd" — полный привод).
9. enginelocation — расположение двигателя в автомобиле (например, "front" — спереди).
10. wheelbase — длина базы колес автомобиля (измеряется в дюймах).
11. enginesize — объем двигателя в кубических дюймах.
12. fuelsystem — система подачи топлива (например, "mpfi" — многоточечный впрыск топлива).
13. boreratio — диаметр цилиндра двигателя (в дюймах).
14. stroke — ход поршня (в дюймах).
15. compressionratio — степень сжатия двигателя.
16. horsepower — мощность двигателя в лошадиных силах.
17. peakrpm — максимальные обороты двигателя (в оборотах в минуту).
18. citympg — расход топлива в городе (миль на галлон).
19. highwaympg — расход топлива на шоссе (миль на галлон).
20. price — цена автомобиля (в долларах).

Каждый автомобиль представлен набором этих признаков, которые характеризуют его технические и эксплуатационные характеристики, а также цену. Эти данные могут быть использованы для анализа, предсказания стоимости автомобилей или изучения взаимосвязей между различными характеристиками.

1. Проработка и анализ данных
На первом этапе работы был выполнен анализ исходных данных. Были исследованы доступные признаки, их распределение, выявлены пропуски и аномалии. Особое внимание было уделено корреляциям между признаками и целевой переменной — стоимостью автомобиля. Это позволило определить важные и бесполезные признаки, а также выявить признаки с высокой корреляцией.

2. Преобразование данных
В процессе работы были выполнены следующие шаги:

- **Удаление высоко коррелированных признаков**:
  Признаки **citympg** (расход топлива в городе) и **highwaympg** (расход топлива за городом) имели высокую корреляцию между собой и низкую корреляцию с целевой переменной (ценой). Было принято решение оставить только один из них — **average_mpg**, представляющий собой среднее значение расхода топлива. Это позволило уменьшить избыточность в данных и улучшить качество модели.
- **Создание новых признаков**:
  Для улучшения качества модели были созданы новые признаки, которые объединяют информацию из высоко коррелированных исходных признаков:
  - Признак **size_ratio**, который отражает отношение размеров автомобиля (ширина × длина) к его весу.
  - Признак **engine_power_index**, индекс мощности двигателя, рассчитываемый как произведение объема двигателя и лошадиных сил, деленное на вес автомобиля. Это позволило учесть как характеристики двигателя, так и вес автомобиля, что важно для оценки его стоимости.
- После создания новых признаков были удалены старые, которые стали избыточными

После создания новых признаков были выявлены высокие корреляции между ними:

- Признаки **price** и **engine_power_index** показали сильную зависимость, что оправдывает сохранение обоих признаков, так как мощность двигателя является важным фактором для оценки стоимости автомобиля.
- Признаки **size_ratio** и **average_mpg** также имели высокую корреляцию. Для улучшения модели был создан новый признак — **efficiency_index**, который объединяет информацию из этих двух признаков, представляя собой соотношение размеров автомобиля к его среднему расходу топлива.

После этого были удалены исходные признаки **size_ratio** и **average_mpg**, так как их информация теперь содержится в новом признаке **efficiency_index**.

3. Обучение модели
Для предсказания стоимости автомобиля была выбрана модель линейной регрессии. Модель была обучена на подготовленных данных, и в процессе обучения использовалась кросс-валидация для повышения точности предсказаний. Признаки были масштабированы, чтобы привести их к единому масштабу, что улучшило качество модели.

4. Оценка качества модели
Для оценки качества предсказаний использовались две метрики:

- **R^2 (коэффициент детерминации)**: показатель того, насколько хорошо модель объясняет вариативность целевой переменной (стоимости автомобиля). Полученное значение R^2 оказалось высоким, что свидетельствует о хорошем качестве модели.
- **MSE (среднеквадратическая ошибка)**: метрика, показывающая, насколько точны предсказания модели. Меньшее значение MSE указывает на более точные предсказания.

Результаты:

- **R^2**: 0.82739
- **MSE**:13626785.65