Выбросы CO2 транспортными средствами

Вопросы:

1. Определите или протестируйте влияние различных переменных на выбросы CO2.
2. Какие особенности наиболее влияют на выбросы CO2?
3. Будет ли какая-либо разница в выбросах CO2 при рассмотрении потребления топлива для города и автомагистралей отдельно и при рассмотрении их взвешенного взаимодействия переменного?

Для определения влияния различных факторов на выбросы CO_2 (CO2 Emissions g/km) можно провести корреляционный анализ и построить регрессионную модель.

1. Корреляционный анализ
На основе представленной матрицы корреляций можно сделать несколько выводов:

- Fuel Consumption Comb (L/100 km) имеет очень высокую положительную корреляцию с выбросами CO_2 (0.92). Это означает, что автомобили с более высоким расходом топлива выделяют больше CO_2.
- Engine Size (L) и Cylinders также сильно связаны с выбросами CO_2 (0.85 и 0.83 соответственно). Это логично, поскольку большие и мощные двигатели потребляют больше топлива, что приводит к увеличению выбросов.
- Fuel Consumption Comb (mpg) имеет отрицательную корреляцию с выбросами CO_2 (-0.91), что ожидаемо, поскольку более высокая топливная эффективность (в милях на галлон) означает меньшее потребление топлива и, соответственно, меньше выбросов CO_2.

2. Какие признаки наиболее влияют на выбросы CO_2?
Наиболее значимые факторы, влияющие на выбросы CO_2, по убыванию корреляции:

1. Fuel Consumption Comb (L/100 km) (0.92) – главный фактор, так как выбросы CO_2 прямо зависят от расхода топлива.
2. Fuel Consumption City (L/100 km) (0.92) и Fuel Consumption Hwy (L/100 km) (0.88) – оба фактора важны, но менее значимы, чем их объединенный показатель.
3. Engine Size (L) (0.85) – большие двигатели вызывают больше выбросов.
4. Cylinders (0.83) – больше цилиндров означает более высокий расход топлива.

3. Разница между анализом потребления топлива в городе, на шоссе и их взвешенной комбинацией

- Городской и шоссейный расход топлива по отдельности сильно связаны с выбросами CO_2, но они взаимосвязаны друг с другом (корреляция 0.95).
- Использование взвешенного среднего ("Fuel Consumption Comb") более логично, так как это обобщенный показатель, отражающий общее топливопотребление в реальных условиях эксплуатации.
- Если рассматривать городской и шоссейный расход топлива по отдельности, возможны небольшие различия в оценках выбросов CO_2, но комбинированный показатель даст более точное представление о реальной картине.

Вывод:

- Главный фактор выбросов CO_2 – общий расход топлива (Fuel Consumption Comb, L/100 km).
- Потребление топлива в городе и на шоссе также влияет на выбросы CO_2, но они сильно коррелируют между собой, поэтому лучше использовать комбинированный показатель.
- Размер двигателя и количество цилиндров также играют важную роль, но через их влияние на потребление топлива.

1. Анализ данных
На начальном этапе был проведен анализ данных, включающий рассмотрение основных характеристик набора данных. Исходные данные содержали как числовые, так и категориальные признаки, среди которых:
- Числовые: размеры автомобиля, мощность двигателя, расход топлива, выбросы CO_2 и другие параметры.
- Категориальные: тип топлива, тип кузова, расположение двигателя, производитель и модель автомобиля.

3. Обучение моделей
Для предсказания стоимости автомобиля использовались две модели:
- Линейная регрессия
- Random Forest Regressor

Модели были обучены на предварительно обработанных данных. Для оценки качества предсказаний использовались метрики:
- Среднеквадратичная ошибка (MSE)
- Коэффициент детерминации (R^2)

4. Результаты и сравнение моделей

Модель	MSE	R^2
Линейная регрессия	0.09	0.91
Random Forest	0.01	0.99

- Линейная регрессия показала хороший уровень предсказания, объясняя 91% дисперсии целевой переменной.
- Random Forest продемонстрировал практически идеальное предсказание ($R^2 = 0.99$), что может свидетельствовать о переобучении.

5. Ответы на вопросы
1. Какие переменные влияют на выбросы CO_2?
- Основные влияющие факторы: размер двигателя, мощность, расход топлива и количество цилиндров. Они имеют высокую корреляцию с выбросами CO_2.

2. Какие признаки оказывают наибольшее влияние на выбросы CO_2?

- `engine_power_index`, `fuel consumption (L/100km)`, `engine size`, `cylinders`. Высокая корреляция указывает, что эти переменные значимо связаны с выбросами.

3. Есть ли разница между потреблением топлива в городе и на трассе?
- Да, но данные показывают, что они сильно коррелируют друг с другом. Поэтому был создан агрегированный признак `average_mpg`.

6. Выводы
- Random Forest показал высокую точность, но требует проверки на переобучение (возможно, стоит ограничить глубину деревьев или использовать меньше деревьев в ансамбле).
- Линейная регрессия демонстрирует приемлемый уровень предсказания, но хуже справляется с нелинейными зависимостями.
- Создание новых признаков и удаление коррелированных улучшило качество моделей.