Подбор гиперпараметров модели

Decision tree

Подготовил: студент группы М8О-307Б-23
Бельский Г.Б.

Почему decision tree?

Decision tree (дерево решений) стоит выбирать, потому что оно обладает рядом важных преимуществ для решения задач классификации и регрессии.

Ключевые преимущества

- Интерпретируемость: Логика дерева решений легко понятна и может быть визуализирована, что позволяет объяснить, почему модель приняла тот или иной результат.
- Простота подготовки данных: Не требует сложной предобработки, нормализации или масштабирования признаков, а также может работать с пропущенными значениями.
- Работа с разными типами данных: Подходит как для числовых, так и для категориальных признаков.

- Быстрое обучение и прогнозирование: Деревья решений быстро строятся и делают предсказания, что важно для реального времени.
- Обнаружение нелинейных зависимостей: Способны моделировать сложные, нелинейные связи между признаками и целевой переменной.

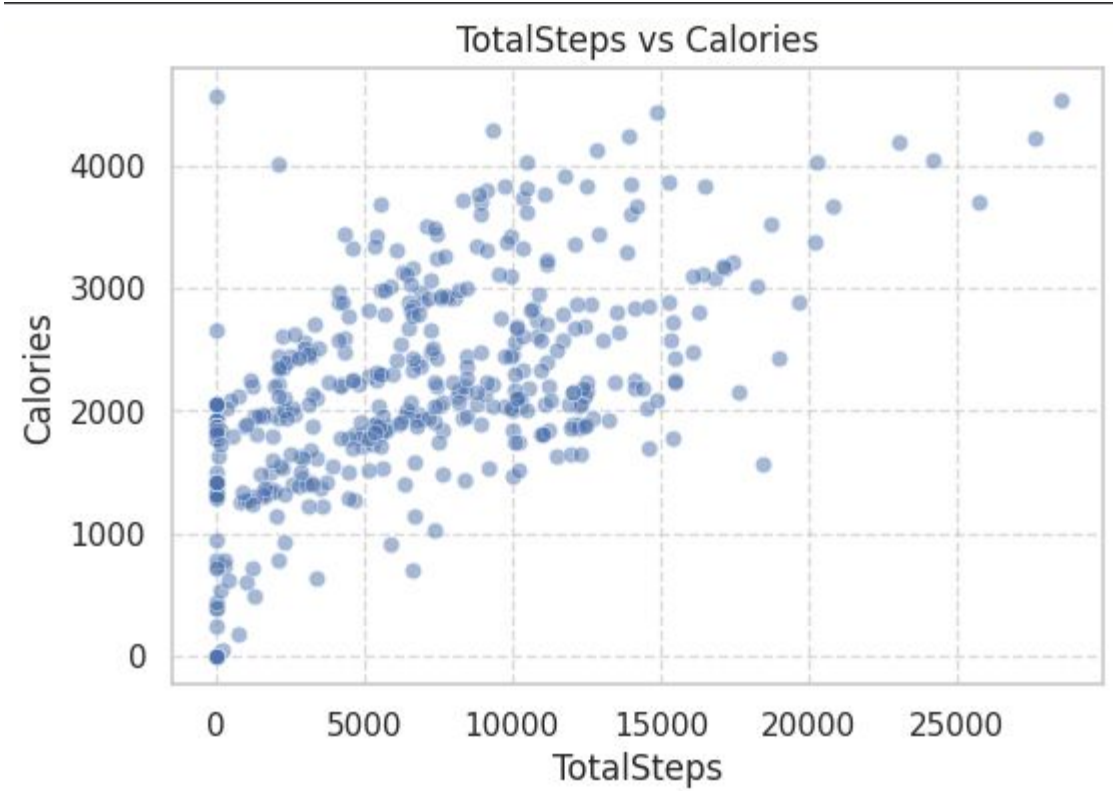
Подготовка датасета

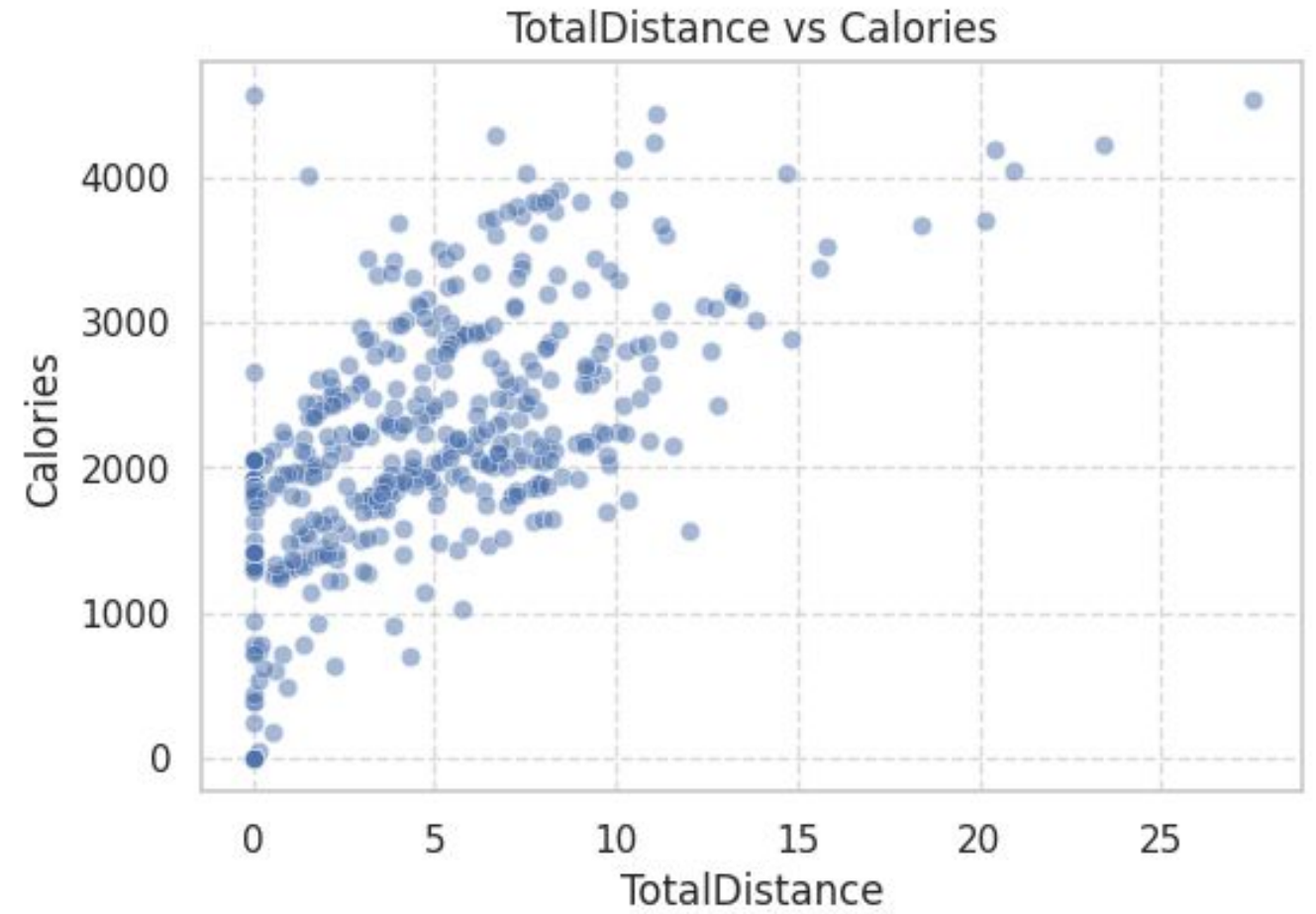
- Деревья решений не требуют сложной предобработки: не нужно нормализовать данные или создавать фиктивные переменные для категориальных признаков.
- Можно работать с пропущенными значениями, но лучше заполнить или удалить их для улучшения качества.
- Для категориальных признаков часто применяется кодирование (например, one-hot), но некоторые реализации (например, scikit-learn) поддерживают категориальные данные напрямую.
- Важно проверить баланс классов (для классификации) и при необходимости использовать стратификацию при разделении выборки.

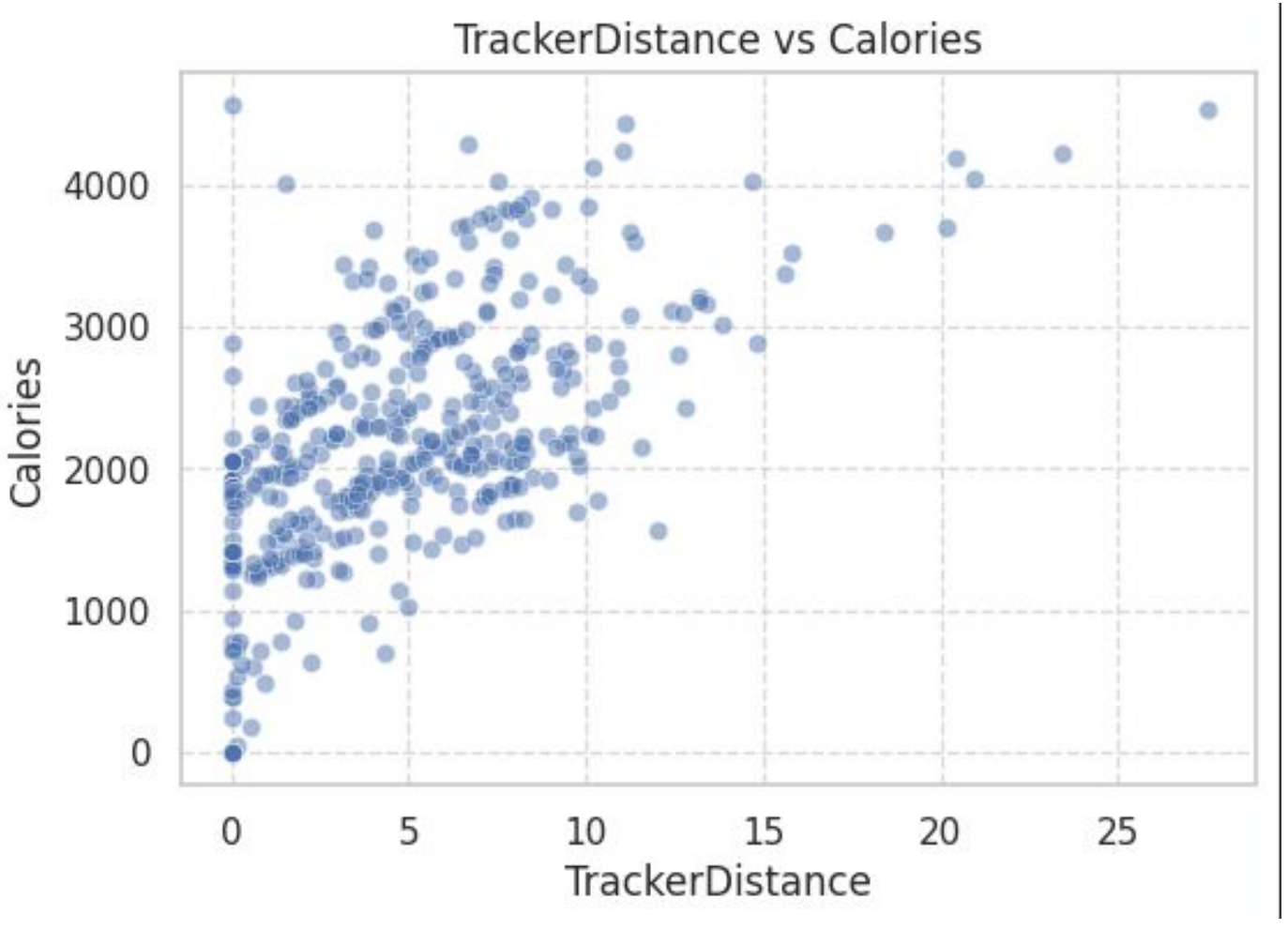
Подбор гиперпараметров

- `max_depth` — максимальная глубина дерева (помогает избежать переобучения).
- `min_samples_split` — минимальное количество объектов для разбиения узла.
- `min_samples_leaf` — минимальное количество объектов в листе.
- `max_features` — количество признаков, учитываемых при поиске оптимального разбиения.

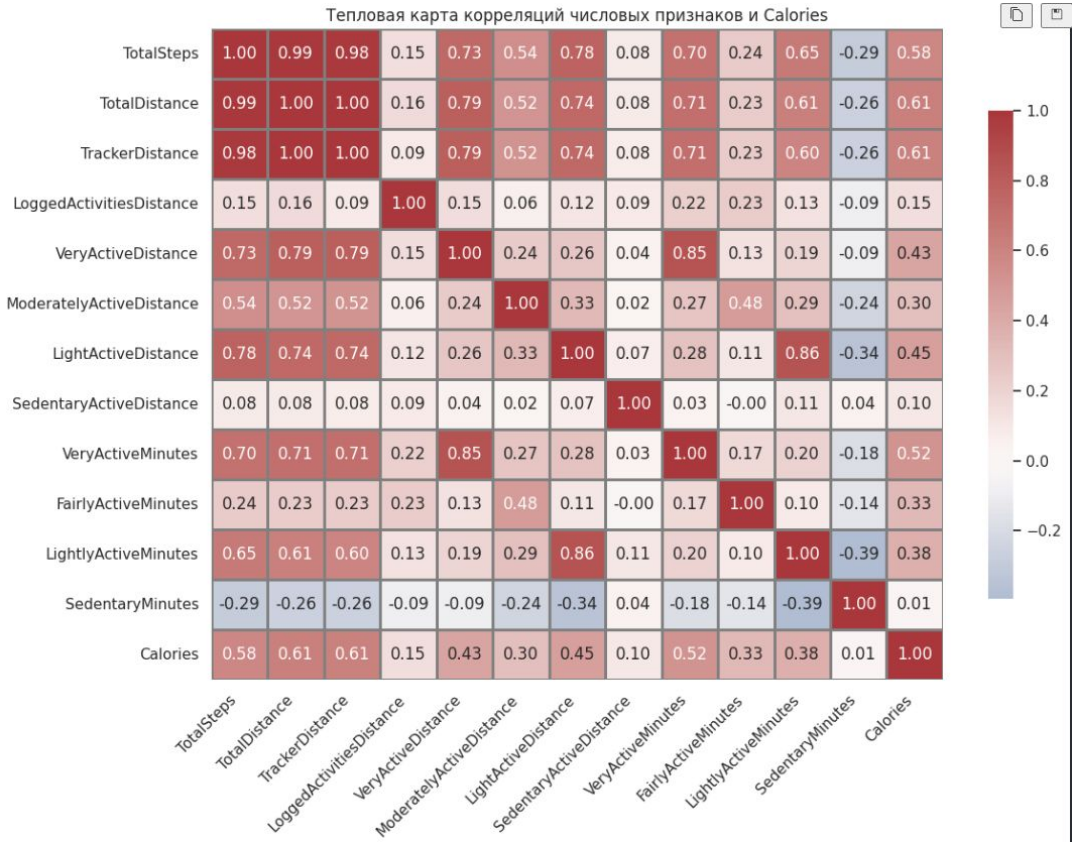
Графики тепловой корреляции параметров с целевой переменной




TotalDistance vs Calories
Calories
4000
3000
2000
1000
0
0
5
10
15
20
25
TotalDistance


TrackerDistance vs Calories
Calories
4000
3000
2000
1000
0
0
5
10
15
20
25
TrackerDistance

Корелляционная матрица



Выводы

Модель получает хорошие базовые результаты (MAE около 304, RMSE 424, R^2 около 0.72), однако после тюнинга гиперпараметров качество может уменьшаться — это может свидетельствовать о переобучении или недостаточной информативности признаков. SHAP показывает, какие признаки наиболее существенно влияют на предсказания дерева, а LIME наглядно объясняет вклад каждого признака в конкретный прогноз на конкретном объекте из тестовой выборки. В целом, ваша лабораторная работа методологически корректна и охватывает все этапы анализа, обучения, подбора гиперпараметров и интерпретации результатов для выбранной задачи регрессии.

Спасибо за внимание!!!