Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

**ПРАКТИЧЕСКАЯ РАБОТА №4**

**Выполнил**:

студент 4 курса ИМРиП

Группы D34101

Ф.И.О. Лю Ифэн

Лю Юе

**Руководитель**:

Чернышева Анастасия Вадимовна

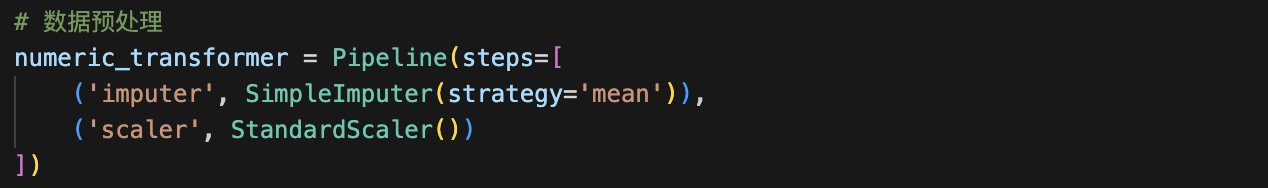
Санкт-Петербург

2024

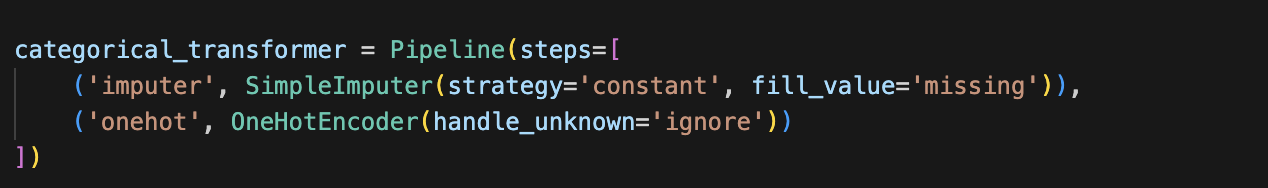
Сначала набор данных делится на набор функций X и целевую переменную y. X содержит все столбцы, кроме «цены продажи», которую мы используем для прогнозирования; и y — целевая переменная, которую мы хотим спрогнозировать, то есть цена продажи автомобиля.

Далее данные подвергаются предварительной обработке, включая обработку числовых и категориальных признаков.

Числовые функции сначала заполняют недостающие значения через SimpleImputer, а стратегия заполнения заключается в использовании среднего значения столбца. Затем с помощью StandardScaler выполняется стандартизация, чтобы гарантировать, что все числовые функции имеют нулевое среднее значение и единичную дисперсию, что помогает модели лучше обучаться и сходиться.



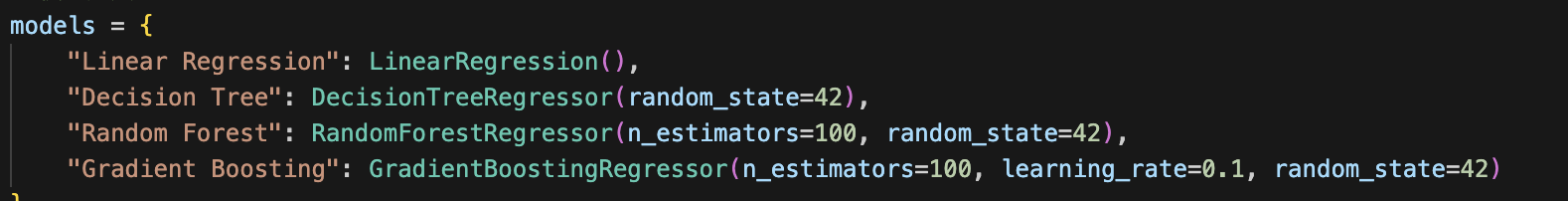
Категориальные функции заполняют пропущенные значения через SimpleImputer, а стратегия заполнения заключается в использовании константы «missing». Затем с помощью OneHotEncoder выполняется горячее кодирование для преобразования категориальных функций в числовой формат, понятный модели машинного обучения.



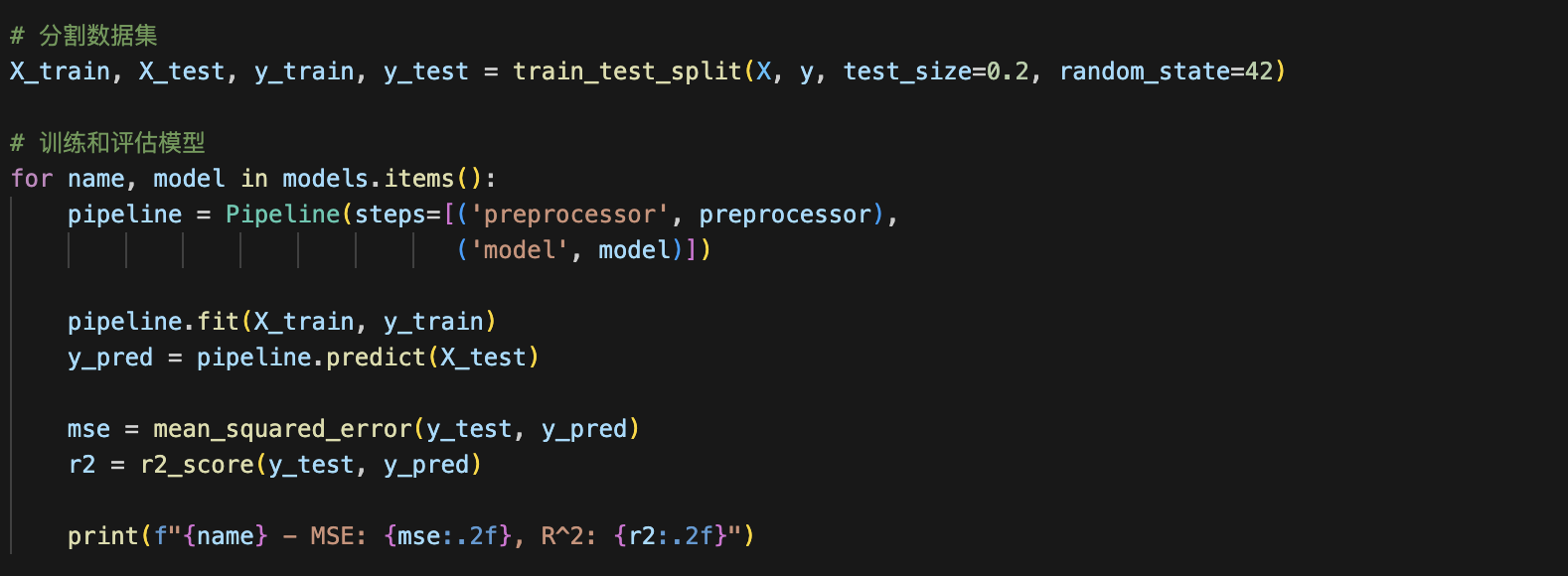
ColumnTransformer используется для применения вышеуказанных преобразований как к числовым, так и к категориальным признакам. Такой подход упрощает процесс предварительной обработки данных и делает код чище.



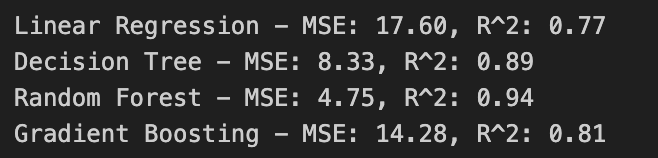
Мы использовали четыре различные модели регрессии: линейную регрессию, регрессию дерева решений, регрессию случайного леса и деревья регрессии с градиентным усилением.



Мы используем train\_test\_split, чтобы разделить набор данных на обучающий набор (80%) и тестовый набор (20%). Затем для каждой модели создается конвейер, содержащий этапы предварительной обработки и саму модель, модель обучается на обучающем наборе, а затем делаются прогнозы на тестовом наборе.



Затем прогностическая эффективность модели оценивалась с использованием среднеквадратической ошибки (MSE) и коэффициента детерминации (R²).



1. Результаты линейной регрессии показывают относительно высокие значения MSE и низкие значения R². Это может быть связано с тем, что взаимосвязь между данными не является полностью линейной или модель не может отразить все взаимосвязи между переменными-предикторами и целевой переменной.

2. Регрессия дерева решений работает лучше: с более низким MSE и более высоким значением R², что показывает, что дерево решений может лучше отражать нелинейные связи между переменными, но может возникнуть риск переобучения.

3. Регрессия случайного леса показала лучшие результаты среди всех моделей с самым низким MSE и самым высоким значением R², что показывает, что путем объединения нескольких деревьев решений для уменьшения переобучения случайный лес может более точно предсказать цену продажи подержанных автомобилей.

4. Производительность дерева регрессии с градиентным усилением находится между линейной регрессией и случайным лесом, показывая лучшую способность прогнозирования, но не так хорошо, как случайный лес.

https://github.com/l769030551/used\_care\_machine\_learning