Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

**ПРАКТИЧЕСКАЯ РАБОТА №5**

**Выполнил**:

студент 4 курса ИМРиП

Группы D34101

Ф.И.О. Лю Ифэн

Лю Юе

**Руководитель**:

Чернышева Анастасия Вадимовна

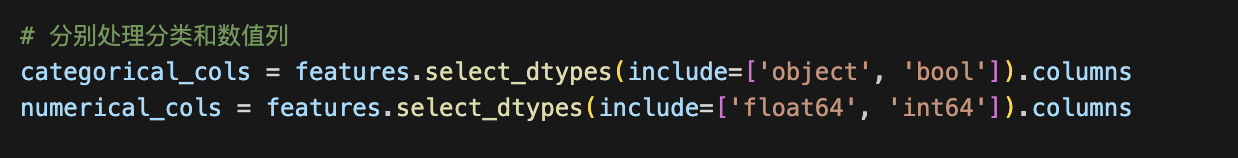
Санкт-Петербург

2024

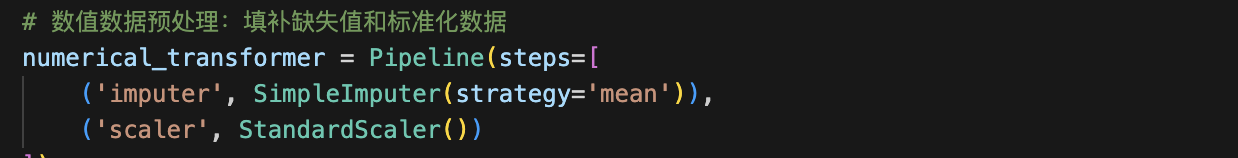
Сначала мы отделяем входные объекты и целевые переменные от набора данных. Здесь «цена продажи» — это целевая переменная, которую мы хотим спрогнозировать, а остальные столбцы используются в качестве входных признаков.



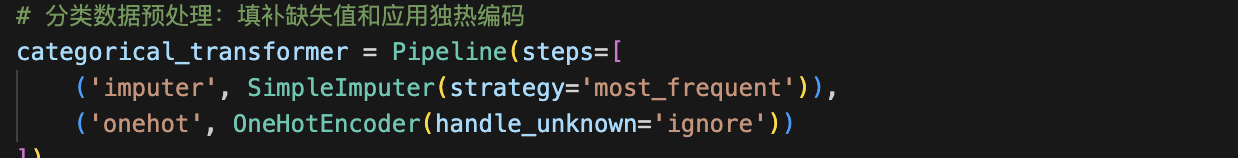
Мы используем метод select\_dtypes, чтобы различать числовые и категориальные функции.



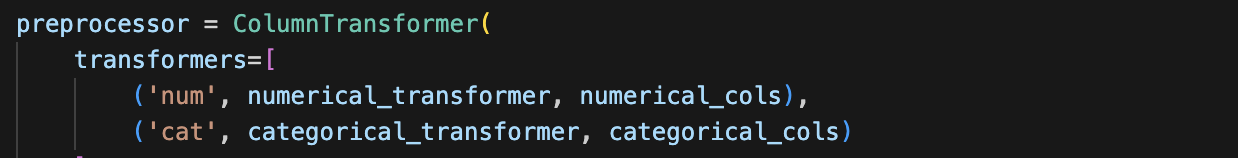
Мы определяем конвейер обработки, который сначала заполняет пропущенные значения, используя среднее значение, а затем нормализует (вычитает среднее значение, делит на стандартное отклонение).



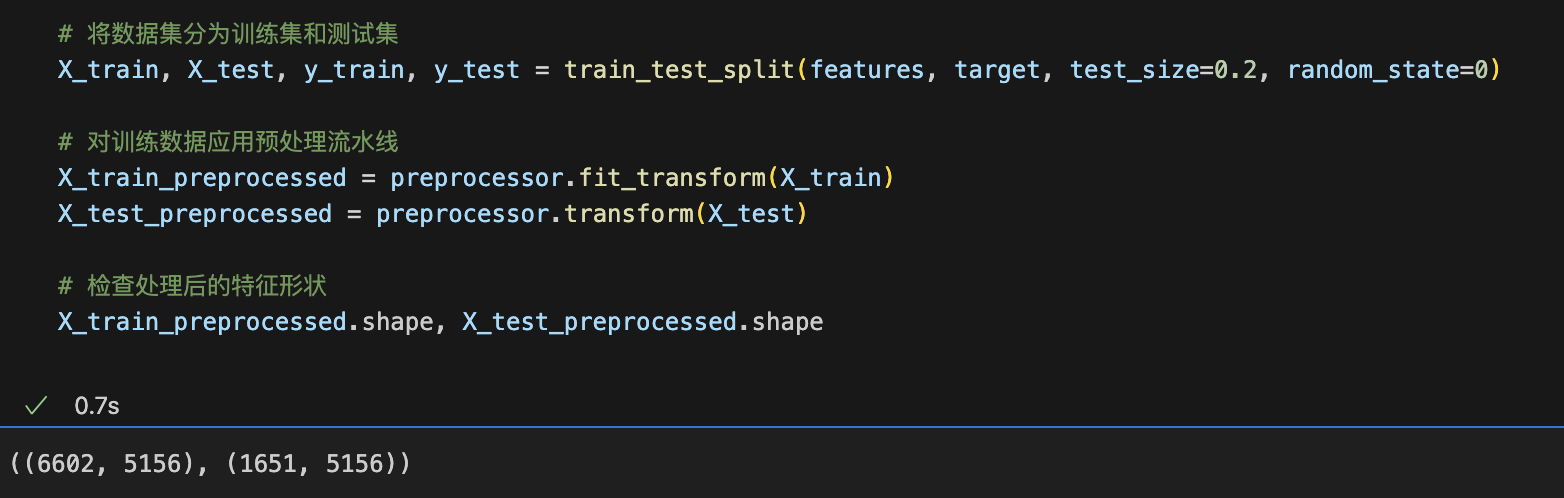
Определите другой конвейер обработки, который сначала заполняет пропущенные значения наиболее часто встречающимися значениями, а затем применяет горячее кодирование.



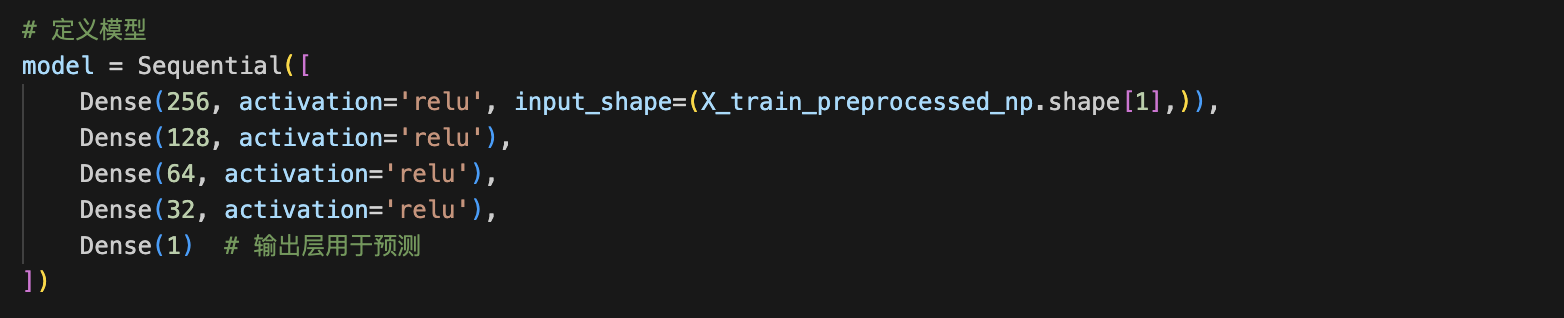
используем ColumnTransformer для объединения этапов предварительной обработки числовых и категориальных данных.



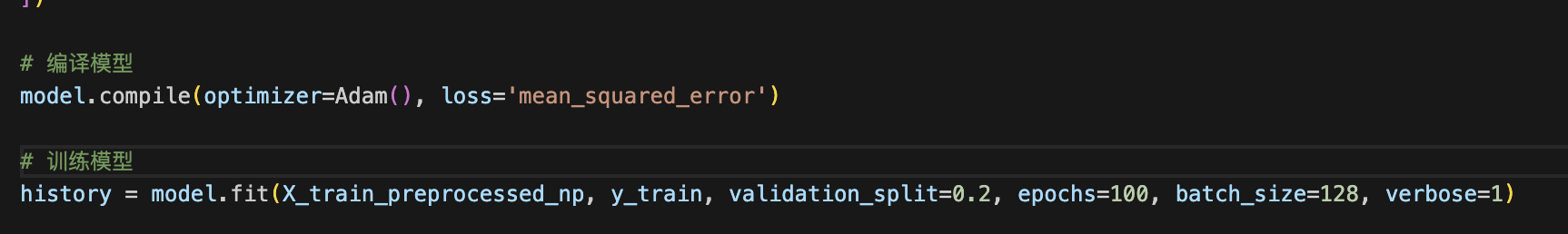
Затем мы случайным образом разделили набор данных на обучающий набор (80%) и тестовый набор (20%).



Здесь применяются ранее определенные шаги предварительной обработки и выводятся размеры обработанных обучающих и тестовых наборов. Результаты, которые вы видите: ((6602, 5156), (1651, 5156)), указывают на то, что обучающий набор содержит 6602 выборки, тестовый набор — 1651 выборка, и каждая выборка содержит 5156 функций.

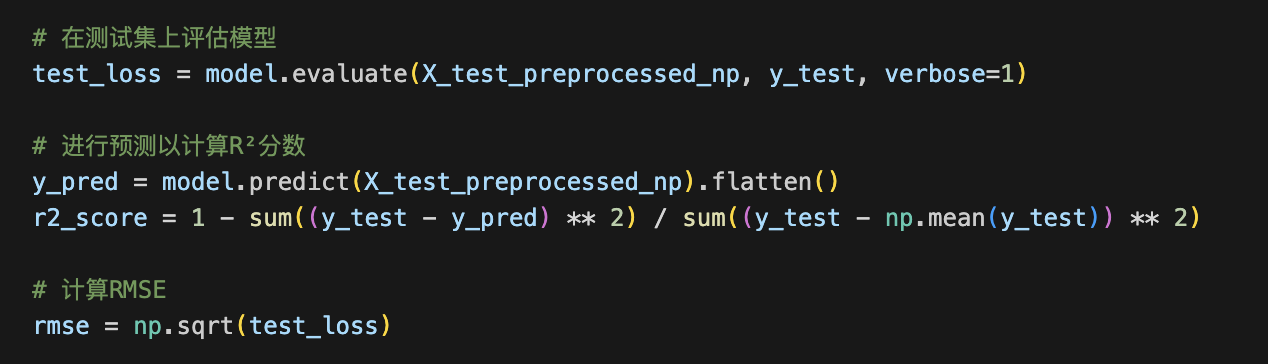


Мы построили последовательную модель, которая содержит несколько плотно связанных слоев, за каждым слоем следует функция активации ReLU, и, наконец, выходной слой с одним узлом используется для прогнозирования отпускных цен.



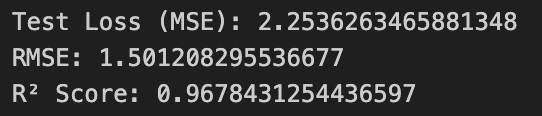
После компиляции модели выберите оптимизатор Адама и среднеквадратическую ошибку в качестве функции потерь.

Для обучения модели мы использовали обработанные данные обучения (X\_train\_preprocessed\_np) и целевые переменные (y\_train). Здесь 20% данных обучения используются в качестве набора проверки для мониторинга производительности во время обучения. Обучение модели проводилось в течение 100 эпох（epochs）, и каждая партия (batch\_size) обрабатывала 128 выборок. Подробный



Мы оцениваем модель на обработанных тестовых данных (X\_test\_preprocessed\_np) и вычисляем потери при тестировании MSE (среднеквадратическая ошибка), RMSE (среднеквадратическая ошибка), показатель R².

Получил следующие результаты：



MSE составляет 2,2536, что означает, что средний квадрат ошибки между прогнозируемым значением модели и фактическим значением составляет 2,2536. Учитывая диапазон и масштаб отпускных цен, это значение можно рассматривать как показатель хорошей производительности модели.

Показатель R² очень близок к 1, что указывает на то, что модель способна точно прогнозировать цены продажи подержанных автомобилей. Высокий показатель R² обычно указывает на хорошее соответствие модели и сильную способность объяснять различия в данных.

Низкое значение RMSE указывает на то, что ошибка прогноза модели относительно невелика, то есть среднее отклонение между прогнозируемым значением модели и фактическим значением невелико.

Заключение

Мы построили высокопроизводительную модель нейронной сети с высокой точностью и надежностью для прогнозирования цен продажи подержанных автомобилей. Эту модель можно использовать в качестве мощного инструмента для предоставления эталонной ценности покупателям и продавцам на рынке торговли подержанными автомобилями.

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