|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по лабораторной работе №3

**«Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Перевощиков Н.Д.

подпись, дата

Проверил:

к.т.н., доц., Гапанюк Ю.Е.

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc161592050)

[2. Задание 3](#_Toc161592051)

[3. Основные характеристики датасета 3](#_Toc161592052)

[4. Листинг 5](#_Toc161592053)

[4.1. Изучение данных 5](#_Toc161592054)

[4.2. Преобразование данных 7](#_Toc161592055)

[4.3. Описательная статистика 8](#_Toc161592056)

[4.4. Предобработка данных 10](#_Toc161592057)

[4.4.1. Пропущенные значения 10](#_Toc161592058)

[4.4.2. Дубликаты 11](#_Toc161592059)

[4.5. Отсев до определенного количества уникальных значений 11](#_Toc161592060)

[4.6. Машинное обучение 13](#_Toc161592061)

[4.6.1. Деление на обучающей и тестовой выборки 13](#_Toc161592062)

[4.7. Кодирование признаков – прямое кодирование (One-Hot Encoding) 13](#_Toc161592063)

[4.8. Обучение модели 14](#_Toc161592064)

[4.8.1. KNeighborsRegressor 14](#_Toc161592065)

[4.9. Итог 15](#_Toc161592066)

[4.9.1. Анализ моделей 15](#_Toc161592067)

[5. Вывод 15](#_Toc161592068)

## Цель лабораторной работы

Изучение способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

## Задание

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите модель ближайших соседей для произвольно заданного гиперпараметра K. Оцените качество модели с помощью подходящих для задачи метрик.
5. Произведите подбор гиперпараметра K с использованием GridSearchCV и RandomizedSearchCV и кросс-валидации, оцените качество оптимальной модели. Используйте не менее двух стратегий кросс-валидации.
6. Сравните метрики качества исходной и оптимальной моделей.

## Основные характеристики датасета

Название датасета: Набор данных о подержанных автомобилях.

Ссылка: https://www.kaggle.com/datasets/shivanink8/used-cars-dataset

**О датасетах**

Этот набор данных содержит информацию о различных аспектах примерно 52 подержанных автомобилей из autotrader.com и carfax.com, включая их марку, модель, год выпуска, пробег, пробег по городу и шоссе, мощность, крутящий момент, мощность двигателя, запас топлива, количество цилиндров, количество посадочных мест, количество владельцев, цену, ссылку, состояние, тип, количество дверей, привод на колеса, тип двигателя, уровни оборотов, запас хода спереди, запас хода для переднего привода, запас хода сзади, запас хода для заднего привода и сервисные записи.

Autotrader - хорошо известный веб-сайт по продаже и покупке подержанных автомобилей. Отчеты Carfax содержат данные, относящиеся к истории автомобиля, такие как количество услуг, типы сервисов, количество владельцев, сообщения о любых авариях и т.д. Таким образом, объединенные данные из обоих этих источников содержат информацию как об автомобиле, так и о его истории.

Имея 24 объясняющие переменные, описывающие различные аспекты подержанных автомобилей в районе Сиэтла, можно предсказать конечную цену автомобилей с помощью регрессии.

**Структура данных**

Id - индекс

brand - марка

model - модель

year – год выпуска

miles - пробег

city mileage – пробег по городу

highway mileage – пробег по шоссе

horsepower - мощность

torque – крутящий момент

engine capacity litre – мощность двигателя

fuel capacity – запас топлива

num cylinder – количество цилиндров

num set – количество посадочных мест

num owners – количество владельцев

price - цена

link - ссылка

condition - состояние

type - тип

doors – количество дверей

wheel drive – привод на колесах

engine type – тип двигателя

speed levels – уровни оборота

front headroom – запас хода спереди

front legroom - передний запас для переднего привода

rear headroom – запас хода сзади

rear legroom - задний запас для заднего привода

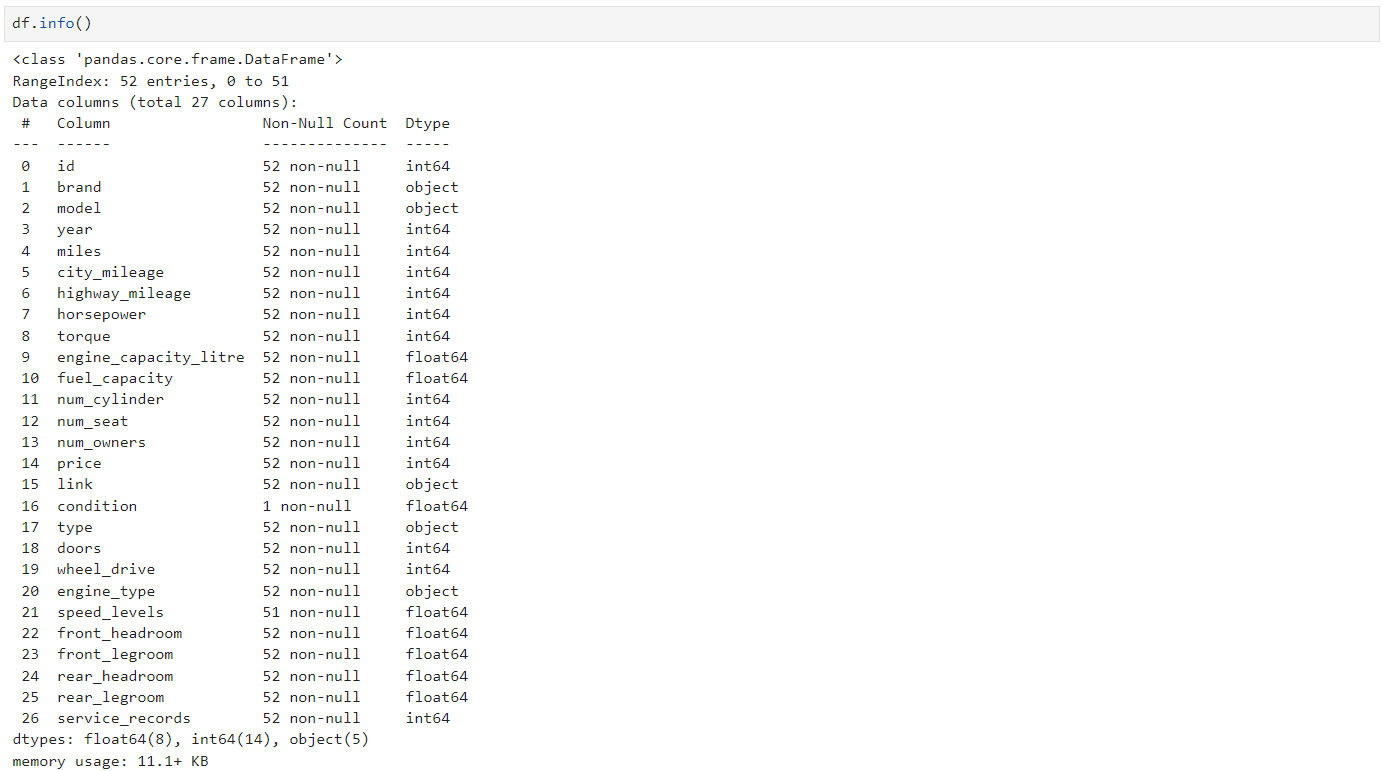
service records - сервисные записи

**Выбор признаков для машинного обучения**

Для машинного обучения выберем целевой признак - стоимость автомобиля. Сопоставим с остальными признаками, а именно, характеристики и конфигурации автомобиля выявляем примерную стоимость автомобиля.

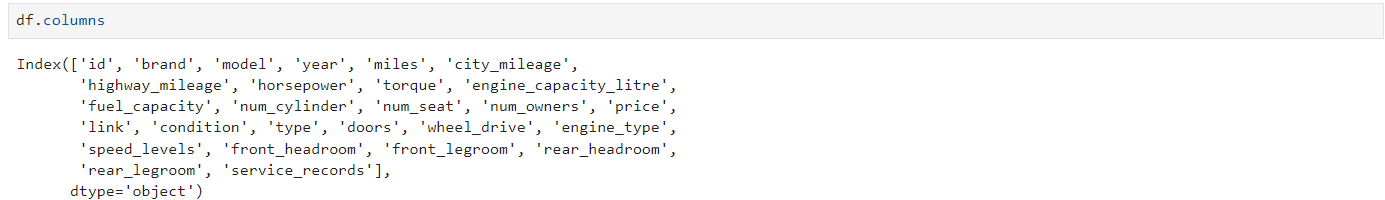
## Листинг

## Изучение данных

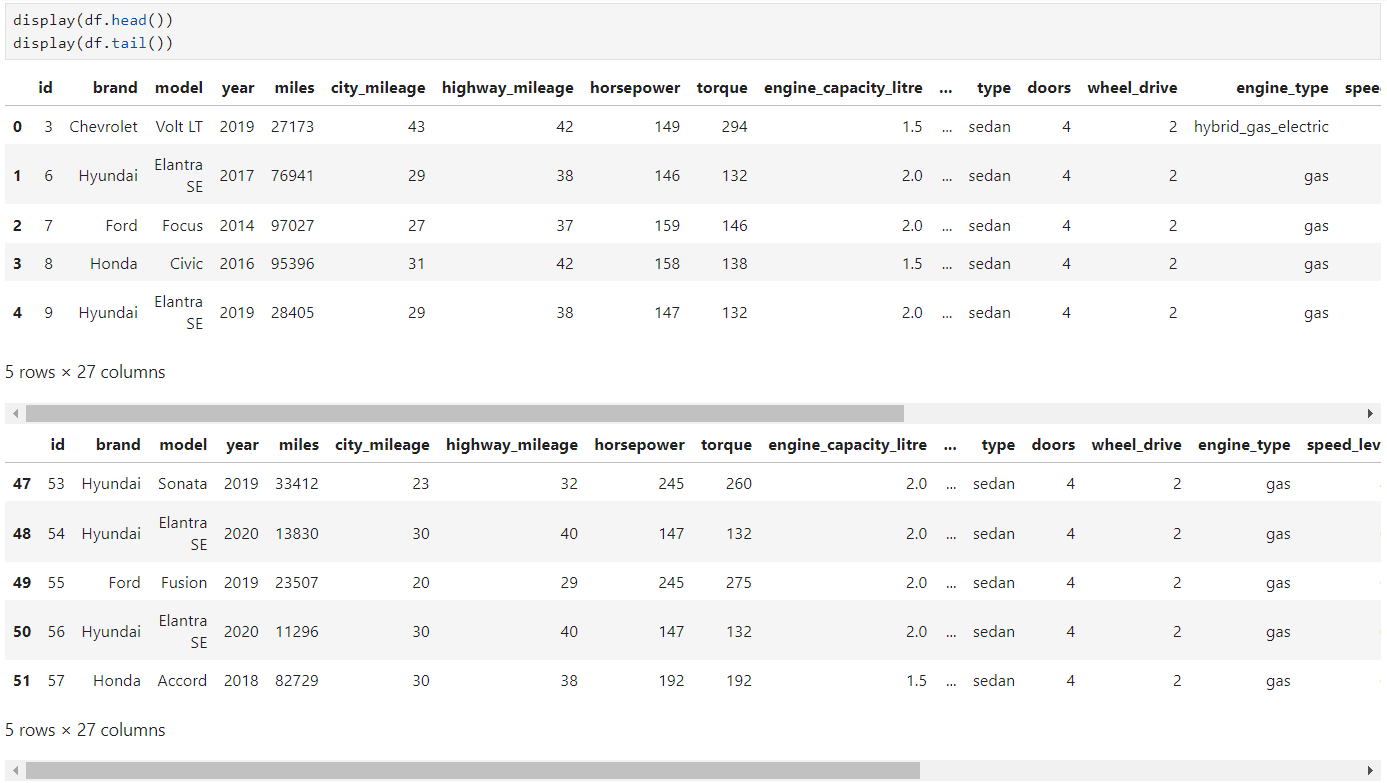


В датасете содержатся 52 строки, имеются 3 различные типы: object, int64 и float64.

Просмотр названий столбцов.



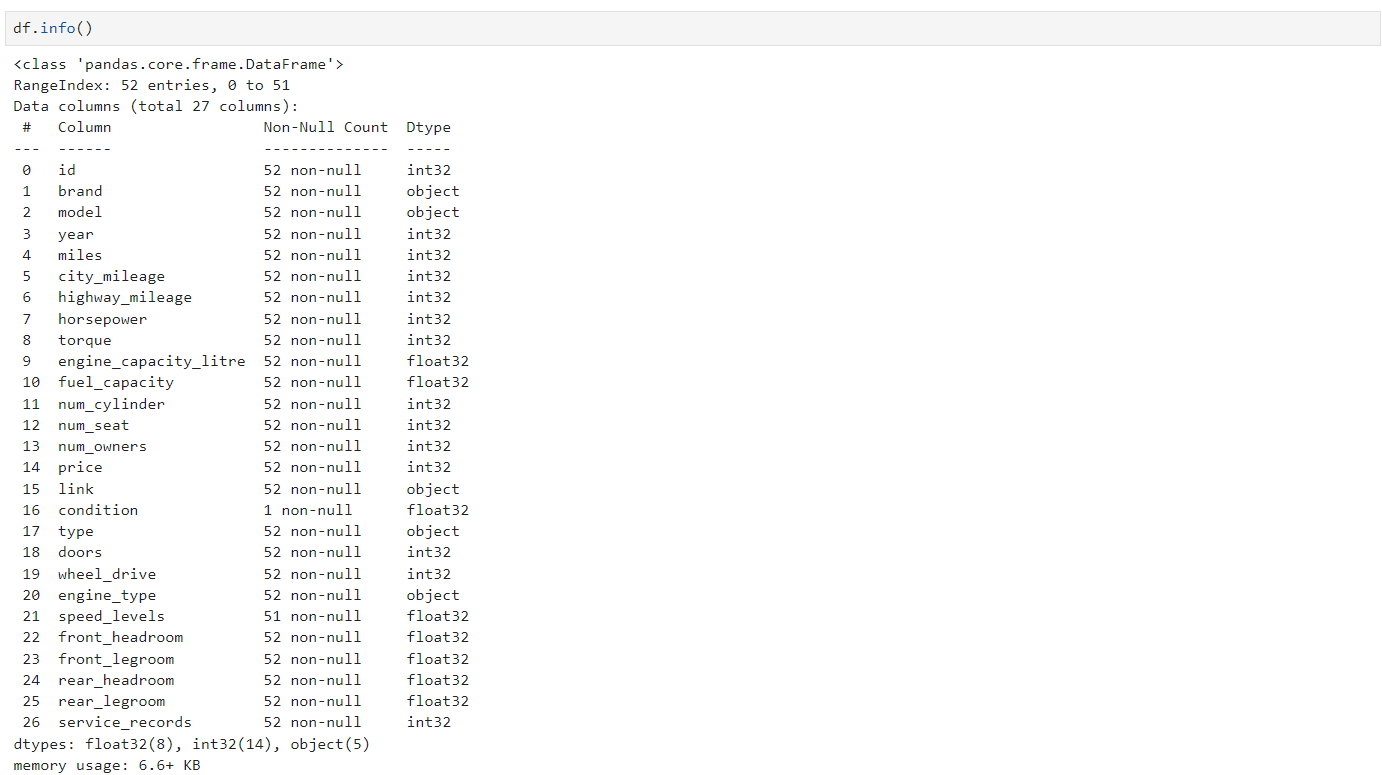
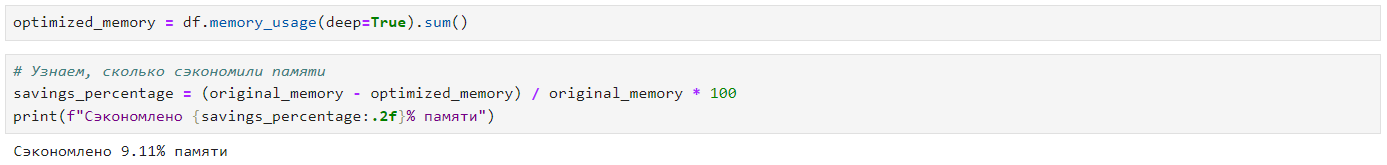
Первые и последние пять строк датасета.



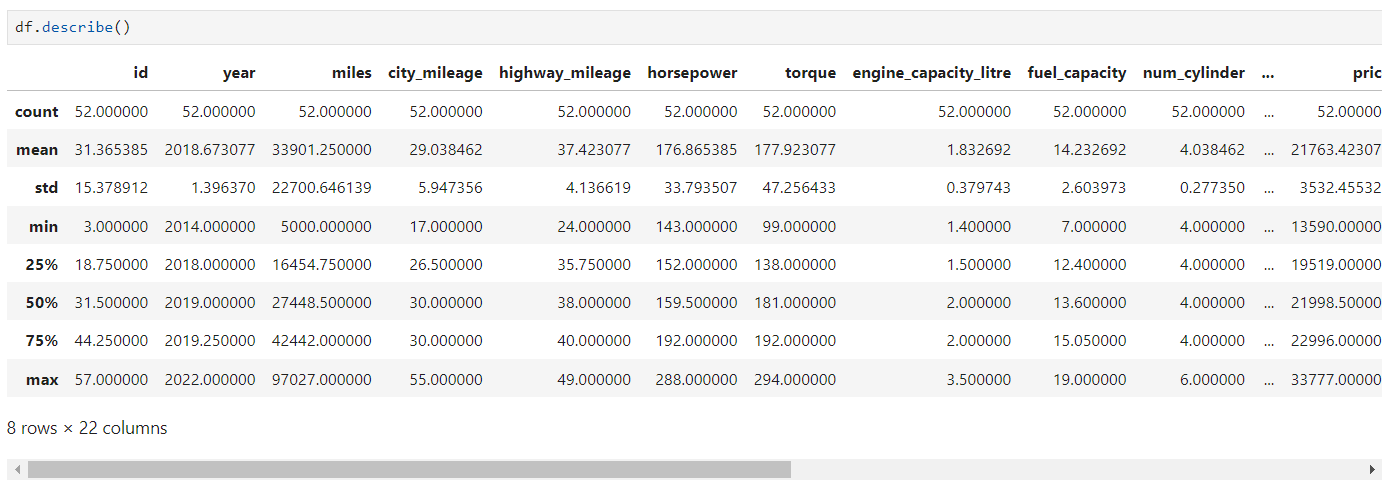
Приведем к нижнему регистру и удаление лишних пробелов в названиях столбцов.



## Преобразование данных



## Описательная статистика



**id**: Это идентификаторы записей или объектов в данных. Каждая строка имеет уникальный идентификатор. Никаких конкретных выводов о данных не делается на основе этого столбца.

**year**: Это год выпуска автомобиля. Данные варьируются от 2014 до 2022 года. Средний год выпуска около 2018.7. Большинство автомобилей в диапазоне от 2018 до 2019 года, что может указывать на то, что эти данные собраны за последние несколько лет.

**miles**: Это количество миль, пройденных автомобилем. Среднее значение составляет около 33901 мили, с большим стандартным отклонением, что может указывать на значительный разброс в пробеге.

**city\_mileage**: Расход топлива в городе, измеренный в милях на галлон. Среднее значение около 29 миль на галлон.

**highway\_mileage**: Расход топлива на шоссе, измеренный в милях на галлон. Среднее значение около 37 миль на галлон.

**horsepower**: Мощность двигателя в лошадиных силах. Среднее значение около 177.9 лошадиных сил.

**torque**: Крутящий момент двигателя. Среднее значение около 177.9, что может указывать на связь между мощностью и крутящим моментом.

**engine\_capacity\_litre**: Объем двигателя в литрах. Среднее значение около 1.83 литра.

**fuel\_capacity**: Вместимость топливного бака. Среднее значение около 14.23 галлонов.

**num\_cylinder**: Количество цилиндров. Среднее значение около 4.

**price**: Цена автомобиля. Среднее значение около 21763 доллара. Минимальная цена составляет 13590 долларов, а максимальная - 33777 долларов.

**condition**: Уровень состояния автомобиля. Все автомобили в данных имеют одинаковый уровень состояния (значение 1.0), что может указывать на единый источник данных или на недостающую информацию.

**doors**: Количество дверей. Все автомобили имеют 4 двери.

**wheel\_drive**: Тип привода колес. Все автомобили имеют одинаковый тип привода колес (значение 4.0), что может указывать на то, что в данных преобладают автомобили с одним типом привода.

**speed\_levels**: Количество скоростей. Среднее значение около 2.15, что может указывать на то, что в основном автомобили имеют две передачи.

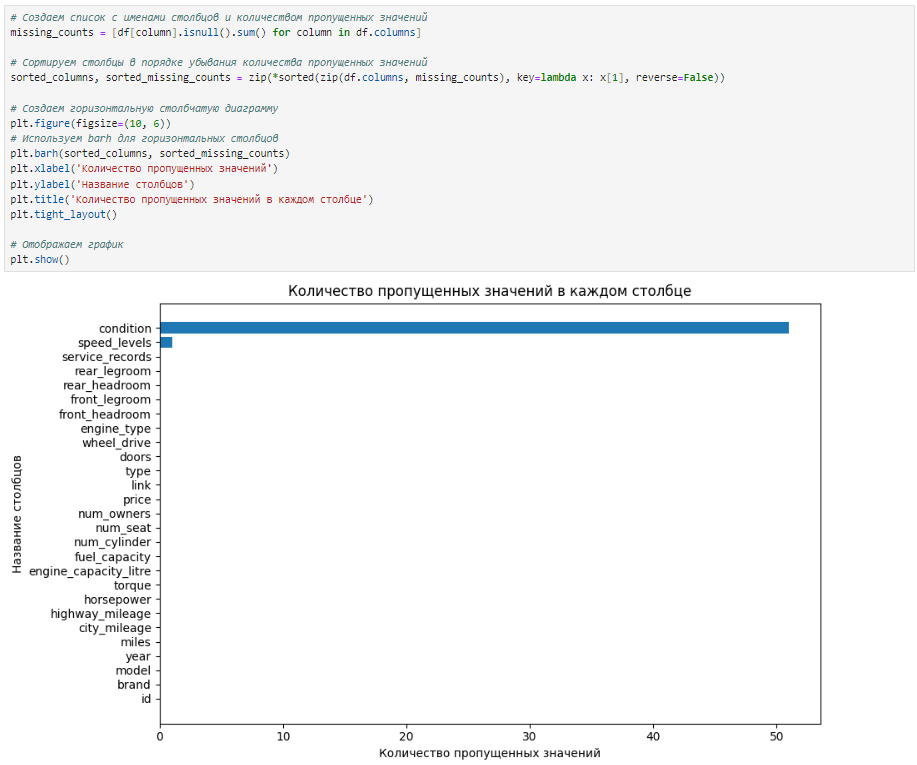
**front\_headroom, front\_legroom, rear\_headroom, rear\_legroom**: Эти столбцы представляют размеры салона автомобиля. Средние значения указывают на средний размер салона для головы и ног для передних и задних пассажиров.

**service\_records**: Количество записей об обслуживании автомобиля. Среднее значение около 8, что может указывать на среднее количество обслуживаний, которое получает автомобиль.

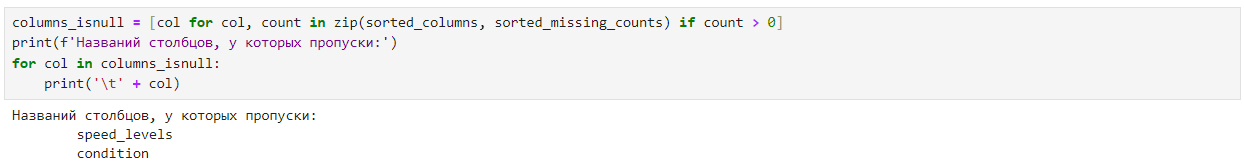
Большинство автомобилей в данных относятся к периоду с 2018 по 2019 год, что может свидетельствовать о том, что данные собраны за последние несколько лет. Пробег автомобилей варьируется от низкого до высокого, что подтверждает разнообразие данных. Средний расход топлива в городе и на шоссе также отражает разнообразие автомобилей в данных. Цены на автомобили также различаются, что может быть связано с их характеристиками и состоянием. Большинство автомобилей имеют 4 двери и один тип привода колес. Средний уровень состояния автомобилей оценивается как одинаковый для всех записей. В среднем автомобили имеют около 2 передач и средний объем двигателя около 1.83 литра. Средний уровень обслуживания автомобилей составляет около 8 записей.

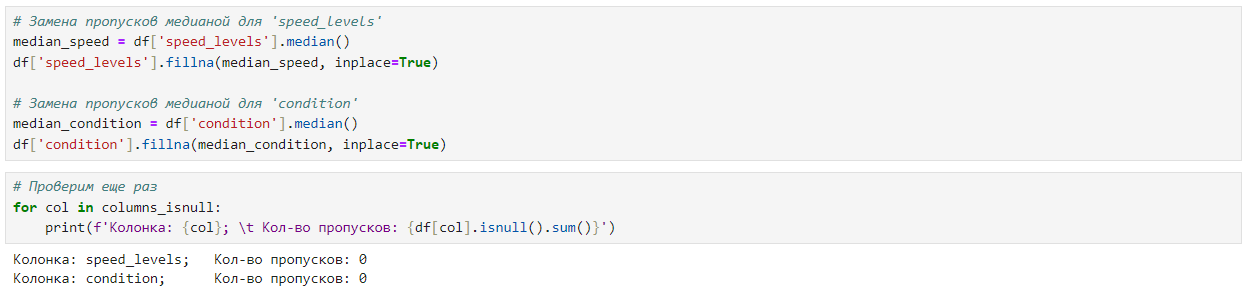
## Предобработка данных

## Пропущенные значения

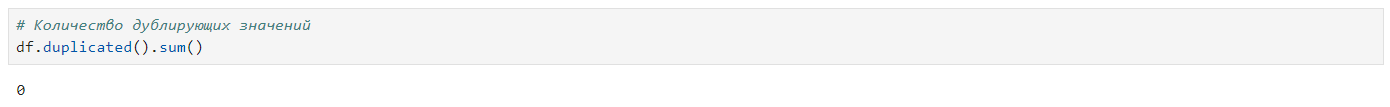


В диаграмме присутствуют большие пропуски в столбцах: состояние и уровни оборота. Заполним пропуски медианными значениями только для численных типов, а остальных типов значения устраняем. Медианными значениями заполняем, потому что они менее чувствительны к выбросам.





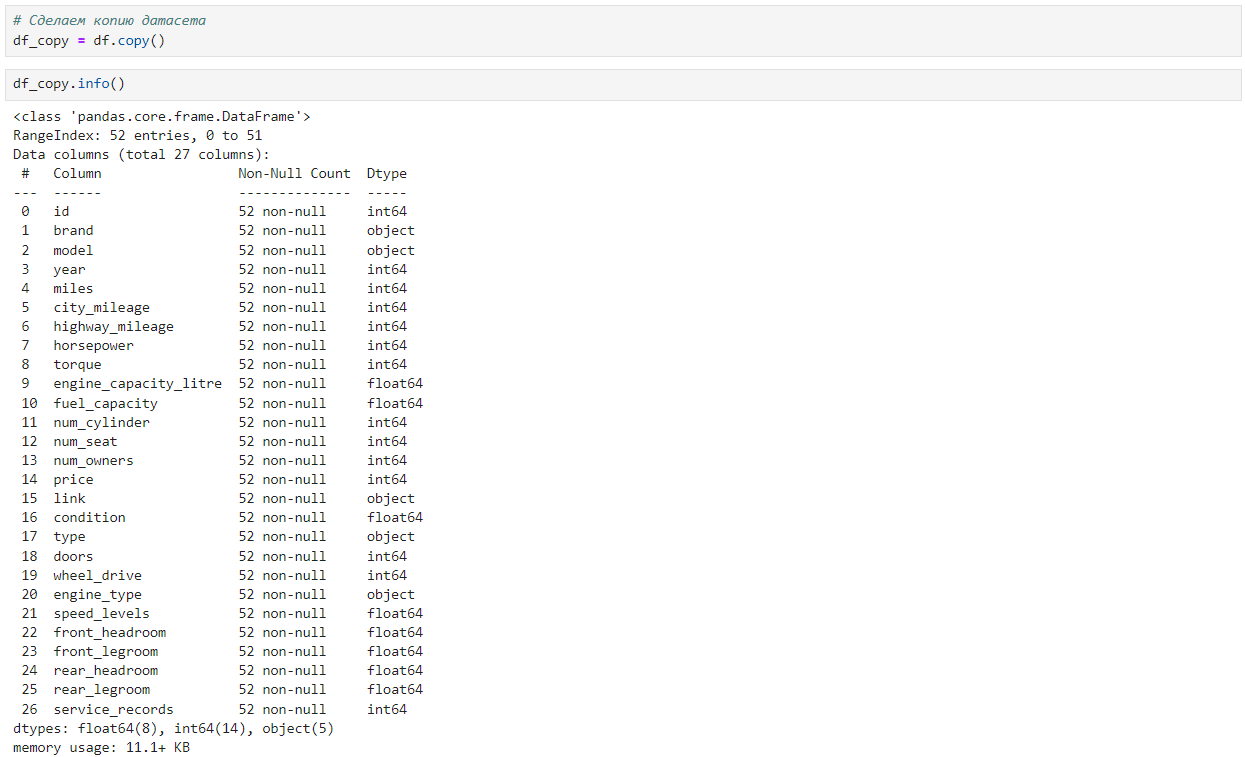
## Дубликаты



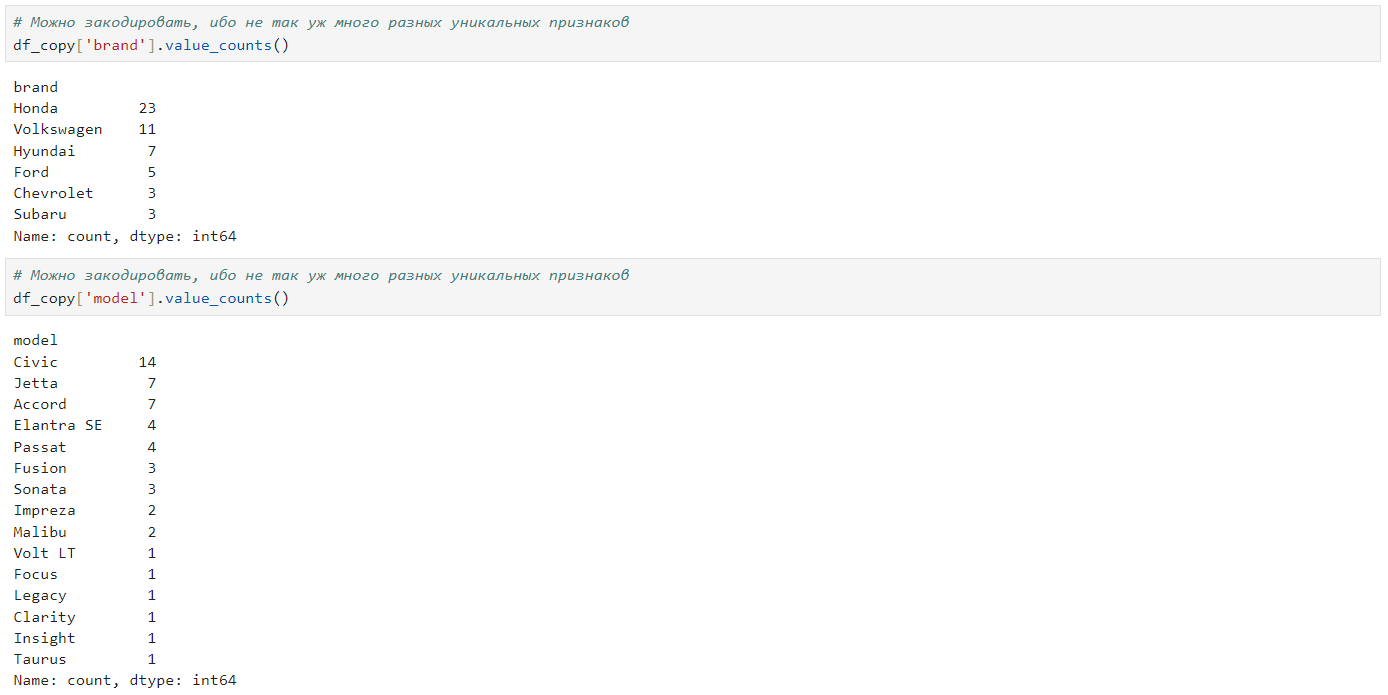
Дубликатов нет, это говорится о том, что датасет был хорошо сделан.

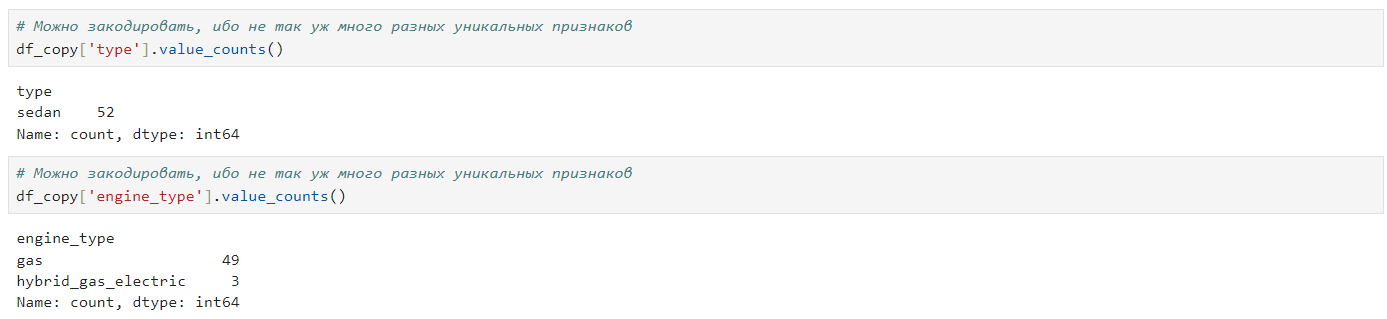
## Отсев до определенного количества уникальных значений

Для кодирования признаков OHE или OH будет черевато, если оставить много уникальных названий, т.к. это приведет к созданию много новых признаков. Поэтому отсеиваем до небольших количеств, так чтобы создали максимум небольших новых закодированных признаков.





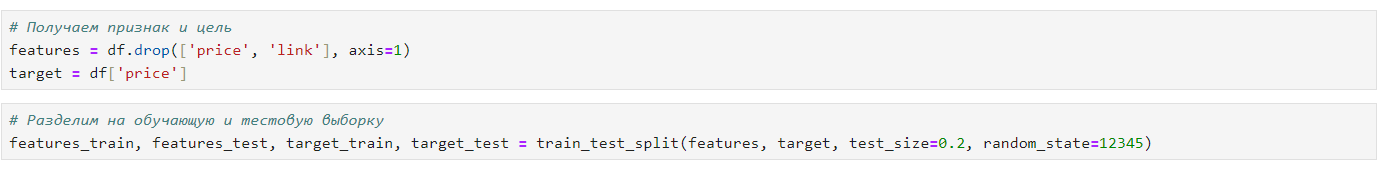




## Машинное обучение

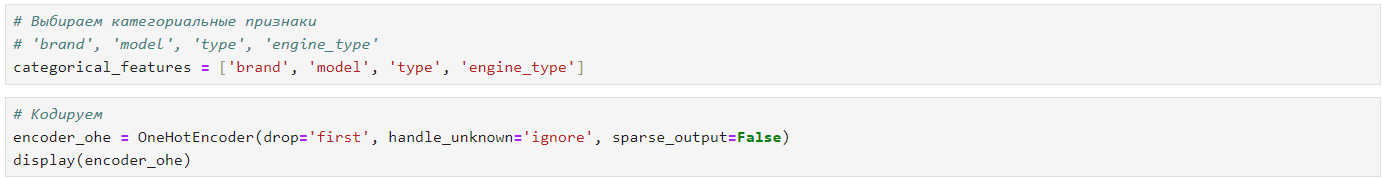


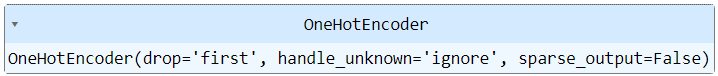
## Деление на обучающей и тестовой выборки

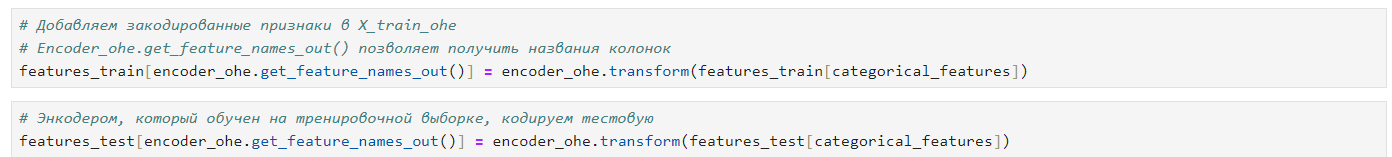


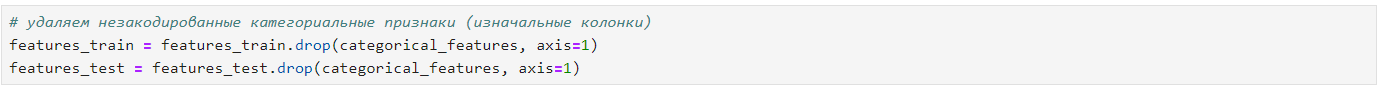
## Кодирование признаков – прямое кодирование (One-Hot Encoding)

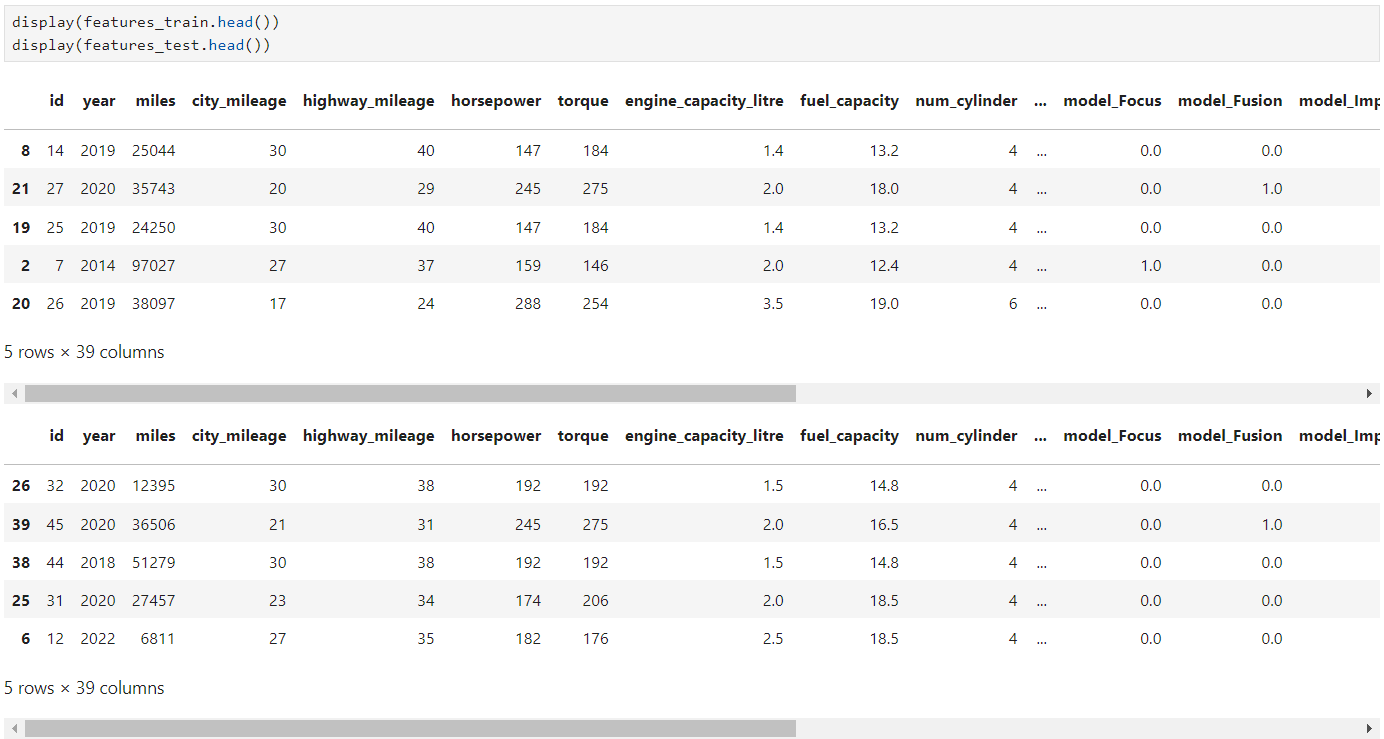
Закодируем признаки: brand, model, type, engine type.

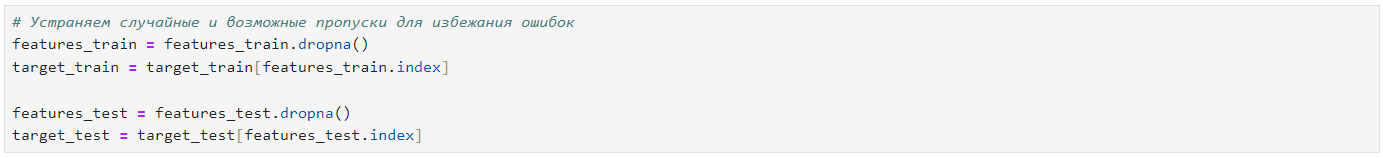






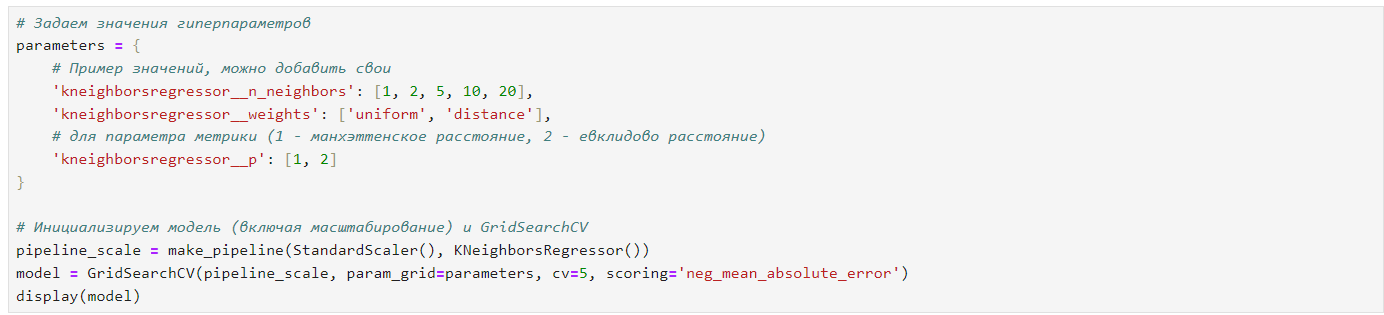
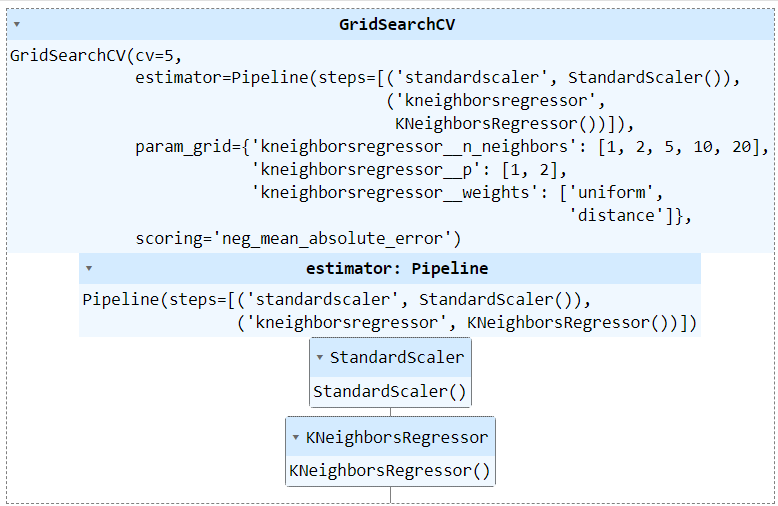


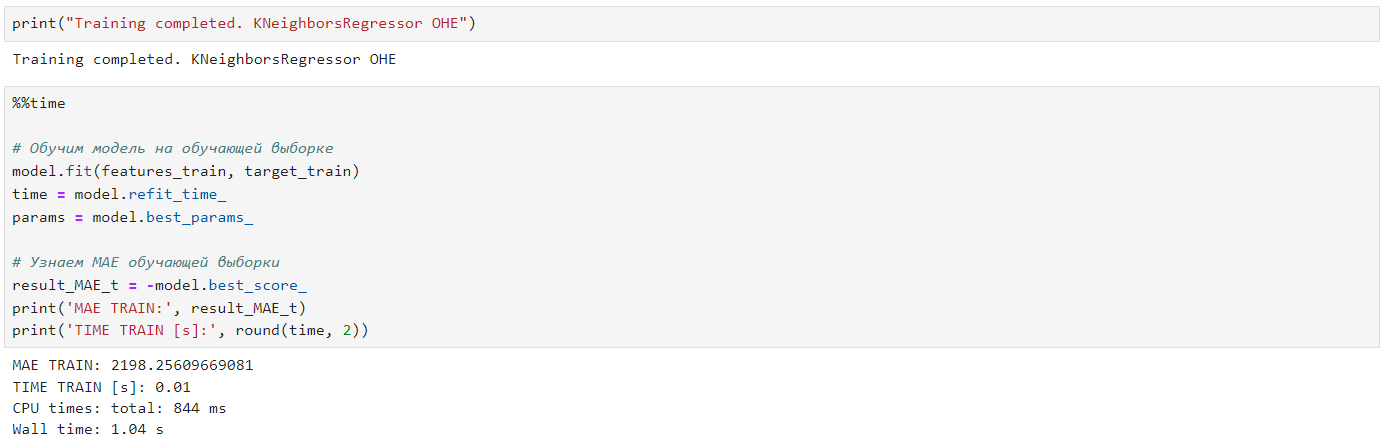




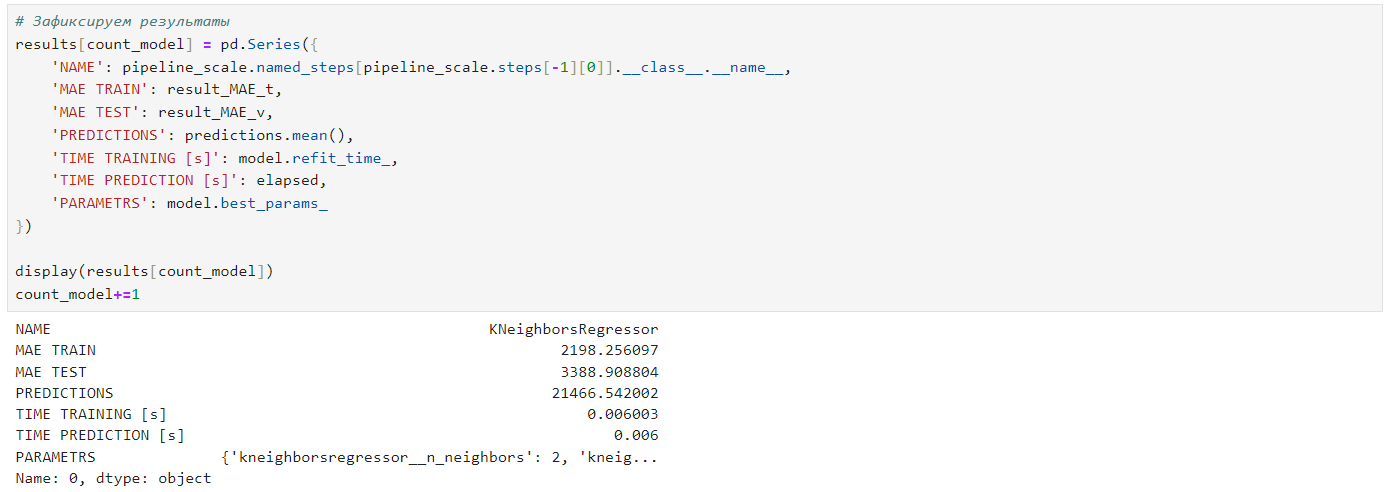
## Обучение модели

## KNeighborsRegressor







## Итог

## Анализ моделей

****

****

## Вывод

В ходе лабораторной работы изучили способы подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.