Моделирование цены подержанных авто

Рабочий проект курса
Data Science

Луцевич Анна

Задача

Строим модель классификации, определяющую категорию цены подержанного автомобиля в зависимости от характеристик транспортного средства

Исходные данные

Работаем с небольшой выборкой из коллекции подержанных автомобилей, выставленных на продажу в США

10 050 строк
27 столбцов

- *url:* URL записи о продаже
- *region:* регион
- *region_url:* URL региона
- *price:* стоимость
- *year:* год выпуска
- *manufacturer:* производитель
- *model:* модель
- *condition:* состояние
- *cylinders:* количество цилиндров
- *fuel:* тип топлива
- *odometer:* количество пройденных миль
- *title_status:* статус
- *transmission:* коробка передач
- *VIN:* *идентификационный номер*
- *drive:* *тип привода*
- *size:* *размер*
- *type:* *кузов*
- *paint_color:* *цвет*
- *image_url:* *URL изображения*
- *description:* *указанное описание*
- *county:* *страна*
- *state:* *штат*
- *lat:* *широта*
- *long:* *долгота*
- *posting_date:* *дата размещения объявления о продаже*
- *price_category:* *категория цены*

7 столбцов числовые данные, остальные – качественные, много пропущенных данных

Порядок действий

- ✓ Проверка данных на состоятельность
- ✓ Выбор признаков, генерация дополнительных
- ✓ Моделирование
- ✓ Оценка модели, выбор лучшей
- ✓ Дополнительная аналитика

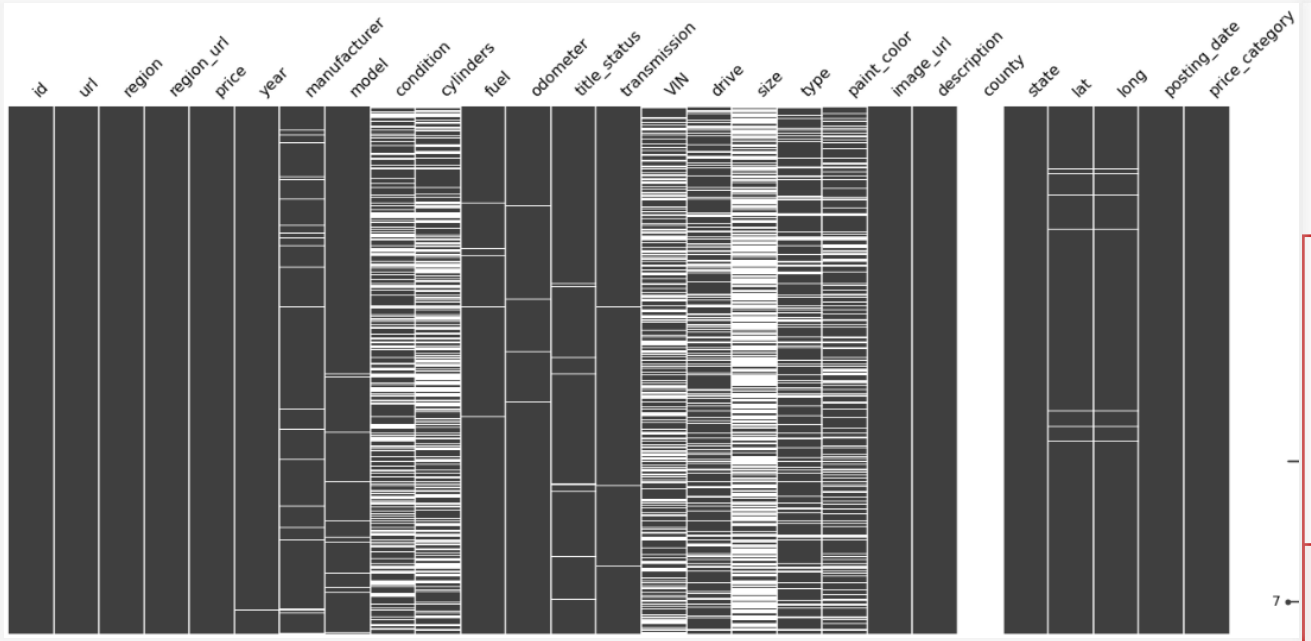
Целевая переменная «price_category»

Имеет равномерное распределение по представленным данным

price_category	col_0 price_category
high	0.349652
low	0.322587
medium	0.327761

EDA

- Удалены дубликаты – выборка сократилась до 10 000 строк
- Из 27 столбцов только 7 заполнены полностью, перед началом обработки нет ни одной строки, заполненной полностью, без столбца “county” – только 8,2% данных полностью заполнены
- Удаляем столбцы с отсутствующей информацией > 20%, полностью заполнены – 91,3% данных
- Удаляем столбцы с отсутствующей информацией > 20%



Процент пропущенных значений:

id	0.00
state	0.00
price	0.00
price_category	0.00
region	0.00
url	0.00
region_url	0.00
description	0.02
image_url	0.02
posting_date	0.02
year	0.36
odometer	0.43
transmission	0.45
fuel	0.63
lat	0.98
long	0.98
model	1.28
title_status	1.66
manufacturer	3.82
type	21.94
paint_color	29.50
drive	30.06
condition	37.66
VIN	39.38
cylinders	40.80
size	71.91
county	100.00

- После отработки пропущенных значений, процент заполнения – 100%, количество строк – 9 868

"manufacturer"	3,82%	заполняем "other"	"title_status"	1,66%	заполняем значением моды
"odometer"	0,43%	заполняем средним	"lat"	0,98%	удаляем строки с пустой широтой
"year"	0,34%	удаляем строки с пустым годом	"long"	0,98%	удаляем строки с пустой долготой
"fuel"	0,62%	заполняем "other"	"transmission"	0,45%	заполняем "other"
"model"	1,27%	заполняем "other"	"posting_date"	0,02%	удаляем строки с пустой датой

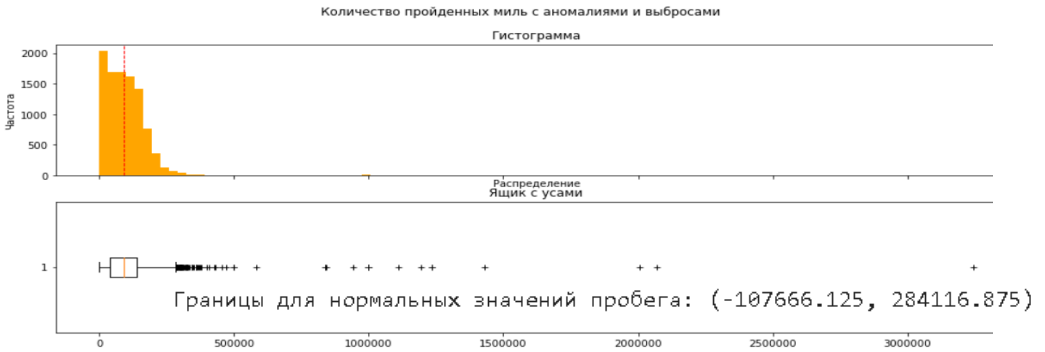
- Изменены типы данных на целочисленные в столбцах – “year” и “odometer”, добавлен столбец даты “date” из столбца “posting_date”

EDA. Оценка и отработка аномалий и выбросов

Исследование данных по пробегу -- "odometer"

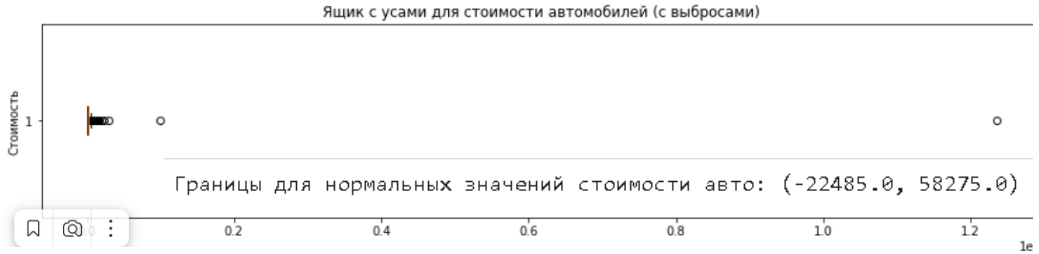
- Явно виды выбросы в данных по пробегу
- По методу квантилей определяем верхние и нижние границы
- Видно, что выбросы в зоне больших значений, заменяем их на верхнюю границу данных пробега

```
Количество значений меньше нижней и больше верхней границ --  90
Процент значений меньше нижней и больше верхней границ --  0.91 %
```



Исследование данных по стоимости авто -- "price"

- Явно виды выбросы в данных по цене, по методу квантилей определяем верхние и нижние границы
- Выбросы только среди макс.значений и они имеют разные характеристики - по году, по модели, удаляем строки с авто самой высокой стоимости




```
Количество значений меньше нижней и больше верхней границ в стоимости авто --  207
Процент значений меньше нижней и больше верхней границ (доля выбросов в стоимости авто)--  2.1 %
```

Исследование данных по году -- "year"

- Явно виды выбросы в данных по годам, выбросы преимущественно в категории низкой цены
- По методу квантилей определяем верхние и нижние границы, т.к. выбросы среди минимальных значений, заполним их значением нижней границы
- Также мало данных по годам 2021-2022 – удалим их из выборки


```
Количество значений меньше нижней и больше верхней границ в годе выпуска автомобиля --  394
Процент значений меньше нижней и больше верхней границ (доля выбросов в годе выпуска автомобиля)--  3.99 %
```

Feature Engineering

✓ 1. Генерация дополнительных фичей

- **odometer/price** — отношение пробега к цене
- **desc_len** — количество символов описания
- **model_in_desc** — наличие модели в описании
- **age_category** — возрастная категория автомобиля
- **model_len** — длина символов в модели
- **model_word_count** — количество слов в модели
- **short_model** — укороченное название модели

✓ 2. Обработка категориальных признаков --

- Преобразование категориальных признаков – manufacturer, fuel, short_model, transmission, region, manufacturer, state, title_status, age_category

✓ 3. Обработка числовых признаков --

- Нормализация количественных признаков – odometer, lat, long, year, odometer/price, desc_len, model_in_desc, model_len, model_word_count

✓ 4. Обработка признака даты --

- Сформированы признаки месяца, дня недели, года, проведена их нормализация

✓ 5. Удаление лишних, исходных столбцов

- Итоговый датафрейм содержит 9 619 строк и 1 494 столбцов

✓ 6. Итоговый датафрейм перед моделированием --

- *url:* URL записи о продаже
- *region_*:* регион
- *x0_*:* вид топлива
- *ls_manufacturer_name: признак* производитель
- *manufacturer_*:* производитель
- *short_model_*:* сокращенная модель авто
- *title_status_*:* статус
- *transmission_*:* коробка передач
- *state_*:* штат
- *age_category_*:* возрастная категория авто
- *std_scaled_odometer*: количество пройденных миль (после стандартизации)
- *year_std*: год выпуска (после стандартизации)
- *lat_std:* широта (после стандартизации)
- *long_std*: долгота (после стандартизации)
- *odometer_price_std:* отношение стоимости к пробегу автомобиля (после стандартизации)
- *desc_len_std*: количество символов в тексте объявления о продаже (после стандартизации)
- *model_in_desc_std:* количество наименований модели автомобиля в тексте объявления о продаже (после стандартизации)
- *model_len_std:* длина наименования автомобиля (после стандартизации)
- *model_world_count_std:* количество слов в наименовании автомобиля (после стандартизации
- *month_std*: месяц размещения объявления о продаже автомобиля (после стандартизации)
- *dayofweek_std:* день недели размещения объявления о продаже автомобиля (после стандартизации)
- *diff_years_std:* кол-во лет между годом производства и годом размещения объявления о продаже (после стандартизации);
- *price:* стоимость
- *price_category*: категория цены

Modelling

- Инициализирована целевая переменная – «price_category», выборка разделена на тренировочный и тестовой сеты в пропорции 70:30

✓ Обучение моделей на выборках

75,7% • логистическая регрессия - точность на тестовом сете

```
TRAIN ACC: 0.85608198425664
TEST ACC: 0.757103257103257
```

75,6% • случайный лес - точность на тестовом сете

```
TRAIN ACC RF: 1.0
TEST ACC RF: 0.7557172557
```

78,8% • многослойный персептрон - точность на тестовом сете

```
TRAIN ACC MLP: 0.97727610
TEST ACC MLP: 0.787941787
```

✓ Результат после кросс-валидации на тренировочной выборке

	Логистическая регрессия	Случайный лес	Многослойный перспептрон
Среднее значение точности модели	76,5779%	75,5679%	79,4593%
Стандартное отклонение точности модели	0,02009	0,01493	0,01440

- Лучший результат по модели многослойного персептрона, выбираем его

✓ Кросс-валидация на тестовой выборке выбранной модели

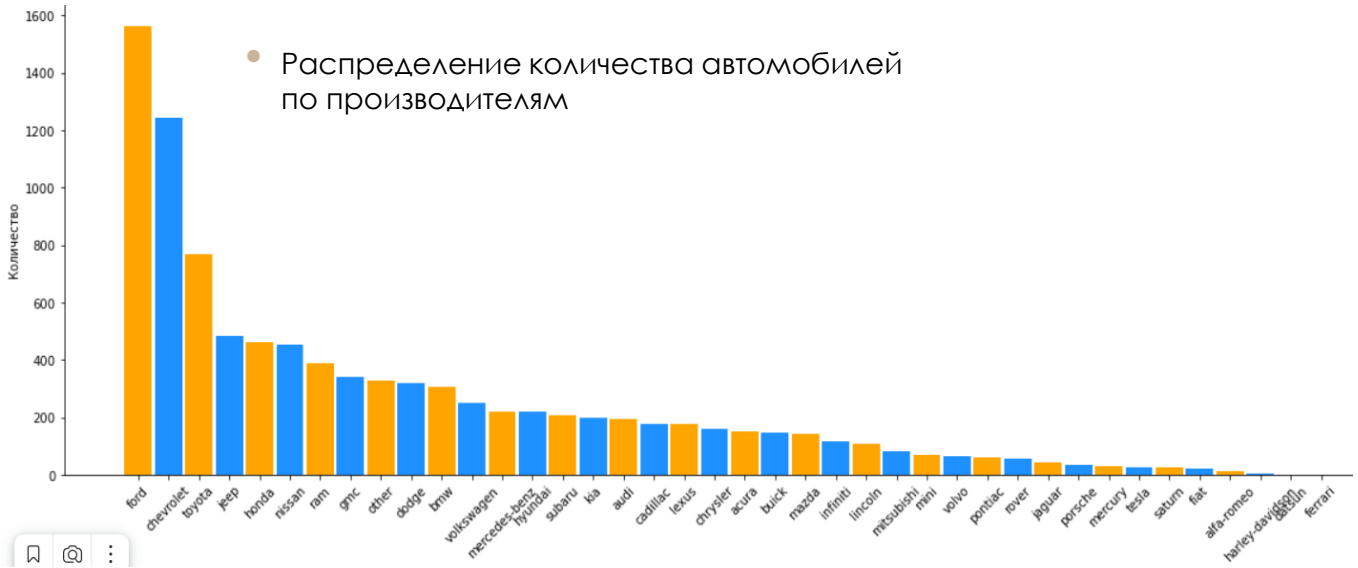
```
Многослойный персептрон на тестовой выборке: среднее значение точности модели -  74.63547882197022 %
Стандартное отклонение точности модели - 0.017751720639139316
```

- по итогам кросс-валидации на тестовой выборке, точность модели ухудшилась незначительно, модель не переобучена и может быть использована¶

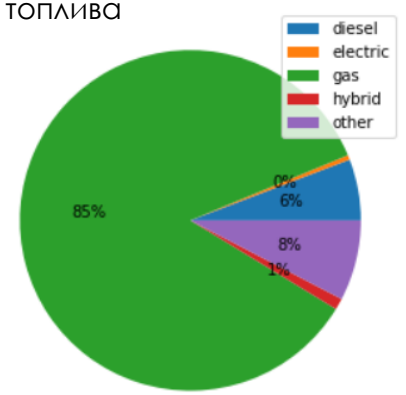
92% • точность выбранной модели на всей выборке

Дополнительные исследования

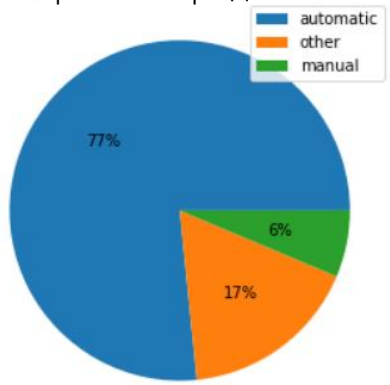
- Распределение количества автомобилей по производителям



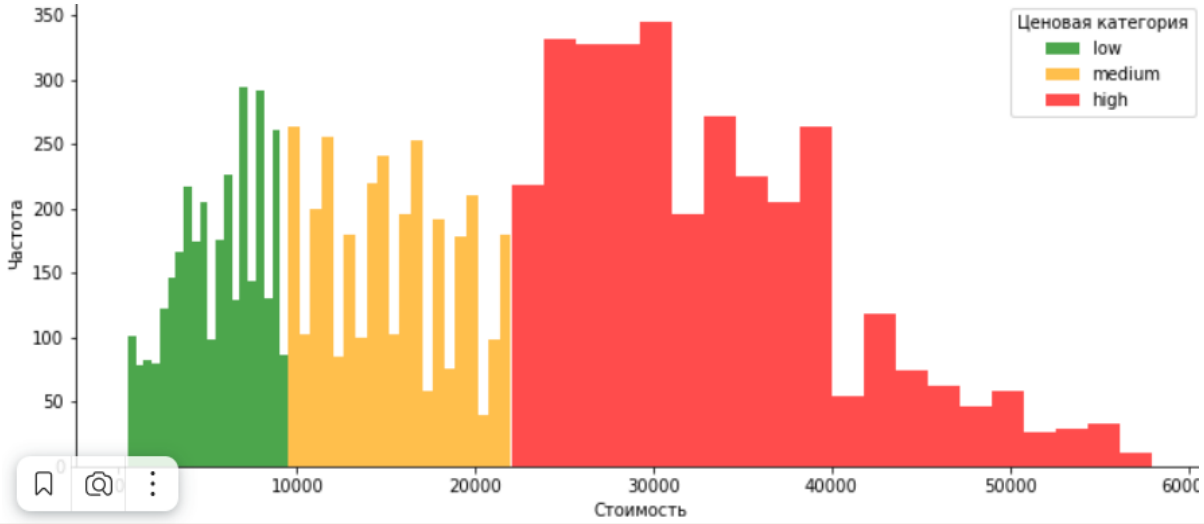
- Распределение типа топлива



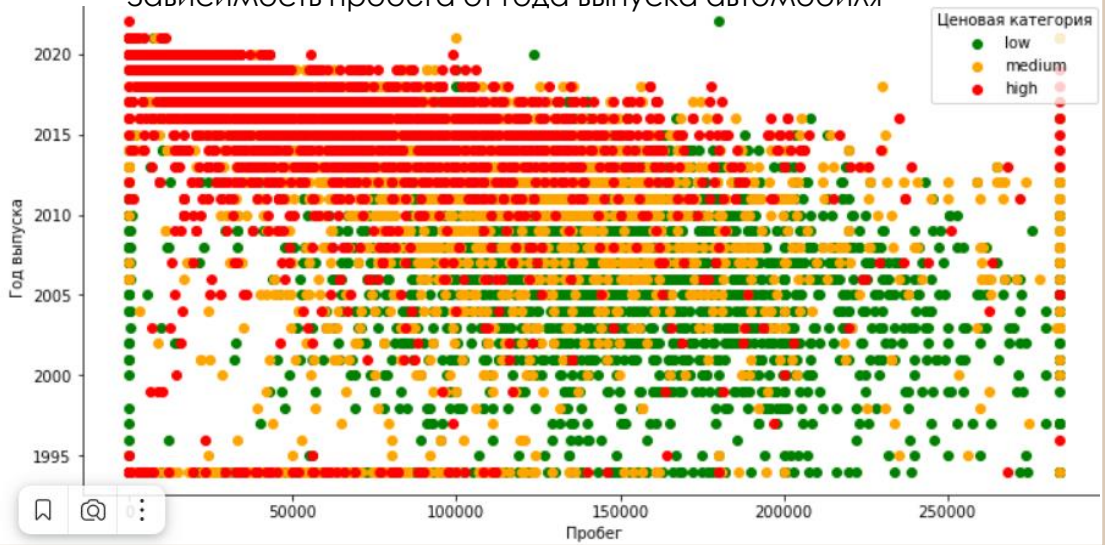
- Распределение типа коробки передач



- Распределение стоимости авто по ценовым категориям

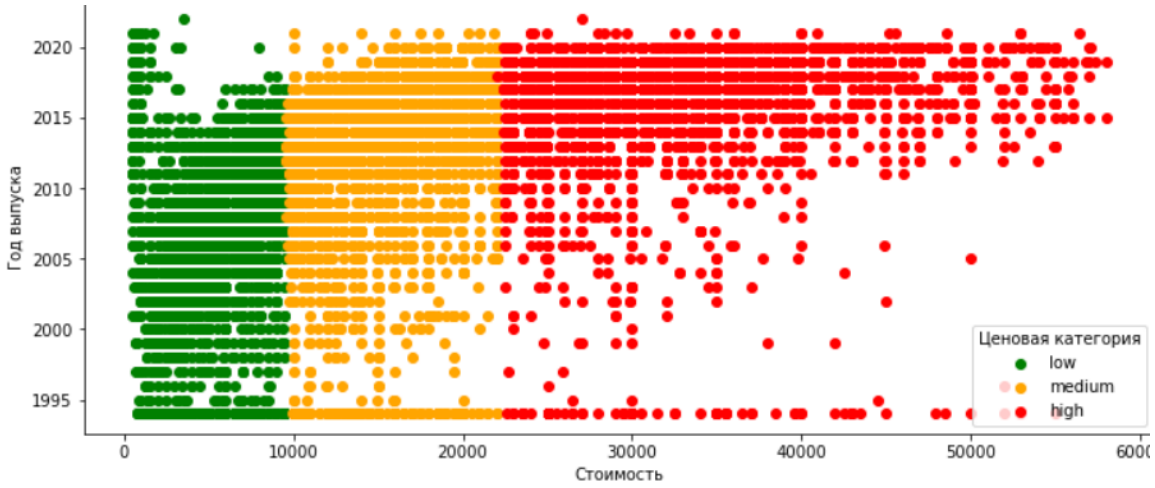


- Зависимость пробега от года выпуска автомобиля



Дополнительные исследования

- Зависимость стоимости авто от года выпуска



- Распределение средней стоимости авто по годам