EDA

1. Типы данных

17 – категориальных признаков

5- числовых признаков

Test и train аналогичны по структуре.

В целом данные без пропусков(за исключением признака «Владение»)

1. Логический контроль категориальных данных(с учетом Boxplot).

|  |  |  |  |
| --- | --- | --- | --- |
| Наименование признака | Значение признака | Значимый для целевой переменной(1 – сильно, 2-средне, 3 - слабо) | Рекомендации |
| bodyType | Тип кузова | 2 | Оцифруем |
| brand | Марка | 2 | Оцифруем |
| color | Цвет | 2 | Оцифруем |
| description | Описание состояния и тд | 2 | Распарсим, применим NLP |
| engineDisplacement | Объем двигателя | 1 | Оцифруем |
| enginePower | Мощность двигателя | 1 | Оцифруем |
| fuelType | Тип топлива | 2 | Оцифруем |
| mileage | пробег | 1 | Очень важный признак |
| modelDate | Дата модели | 1 | Удалим(дублирование корреляции) |
| model\_info | Класс авто | 1 | Преобразовать в числовой. Сгруппировать |
| name | Общее описание | 1 | Дублирует другие признаки: содержит информацию об объеме двигателя , КП, и прочее. Удалить |
| numberOfDoors | Количество дверей | 2 |  |
| productionDate | Дата выпуска авто | 1 | Очень важный признак.  Оцифруем как возраст авто |
| vehicleConfiguration | комплектация | 1 | Дублирует другие признаки: содержит информацию об объеме двигателя , КП, типа кузова и прочее. Удалить |
| vehicleTransmission | Тип КП | 1 | Оцифруем |
| Владельцы | Количество владельцев по ПТС | 1 |  |
| Владение | Время владения | 3 | Удалим - очень много пропусков |
| ПТС | ПТС | 1 | Оцифруем |
| Привод | Привод | 2 | Оцифруем |
| Руль | Руль | 2 | Удалить |

1. Статистический анализ.

Рекомендации – удалить признак «Руль».

1. Числовые переменные

Посмотрели корреляции.

Увидели не нормальность распределения числовых признаков(будем нормализовывать)

1. Оцифровка категориальных данных.

Что будем оцифровывать:

|  |  |
| --- | --- |
| Признак | Как оцифровываем |
| engineDisplacement | Словарь |
| enginePower | Преобразование |
| fuelType | Словарь |
| model\_info | Попробуем перекодировать в следующие классы авто: A, B, C, D, E, S |
| bodyType | Словарь |
| brand | Словарь |
| color | Dummy, все необычные цвета сделаем как other |
| productionDate | Сделаем как возраст авто |
| ПТС | Словарь |
| vehicleTransmission | Словарь |
| Привод | Словарь |

1. Удаление не нужных данных

Удаляем следующие признаки:

«Руль», «Владение», «sell\_id», «modeldate», «name», vehicleConfiguration

FE

Добавим следующие признаки:

Налоговая ставка авто;

Класс авто;

Возраст авто;

Дамми переменные