Что было сделано.

1) EDA, FE в части предобработки табличных данных

2) Построена модель на основе градиентного бустинга с помощью CatBoost

3) Построена модель на основе нейронных сетей(для таблчных данных)

4) Построена multi-input модель для анализа табличных данных и текста (используем лемматизацию)

5) Построена multi-input модель с добавлением нейронной сети обработки изображений

6) Сделан блэндинг градиентного бустинга и нейронной сети (усреднение их предсказаний)

EDA, FE

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 и сделать dummy |
| 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

Добавим следующие признаки:

Налоговая ставка авто;

Класс авто;

Возраст авто;

Дамми переменные

Модель на основе градиентного бустинга с помощью CatBoost.

Была сделана попытка подбора гиперпараметров для CatBoost, очень долго считалось – положительного результата получить не удалось.

После обработки датасета наибольшая оценка получена с добавлением полиномиальных признаков

Метрика(МАРЕ %) – 12,62

Модель на основе нейронных сетей(для табличных данных)

Код полностью взят с бейзлайна, анализ метода не проводил.

Multi-input модель для анализа табличных данных и текста (используем лемматизацию)

|  |  |  |
| --- | --- | --- |
| Признак модели(в отличии от бейзлайна) | Метрика(МАРЕ %) | Комментарии |
| Бейзлайн | 12,37 |  |
| С лемматизацией | 12,41 |  |
| С батч нормализацией | 20 | Не понятно как объяснить резкое падение качества метрики |
| С CLR | 14,51 |  |

Finetuning был сделан дальше, при использовании изображений.

Multi-input модель с добавлением нейронной сети обработки изображений

Добавил побор LR(CLR), дополнительный слой из нейросети EfficientNetB7, батч нормализацию, сделал Finetuning

Метрика(МАРЕ %) - 12,3%

Блэндинг градиентного бустинга и нейронной сети (усреднение их предсказаний)

МАРЕ 12,18%

Итоговый результат:

Наибольшей оценки удалось получить после блэндинга Бустинга и нейросети EfficientNetB3