Проект заключается в создании модели для предсказания стоимости автомобиля. Источник данных для обучения модели - сайт auto.ru

Сбор данных напрямую с auto.ru к сожалению оказался непосильной задачей.

Выручили готовые датасеты выложенные на Kaggle. Удалось раздобыть 4 датасета за 2021 год в дополнение к датасету из Baseline. Соответственно, важной задачей было грамотно склеить их в один общий и объединить с тестовыми данными.

После создания объединенного датасета, был проведен тщательный EDA.

ML включил в себя работу с признаками и генерацию новых признаков, создание 6-ти моделей и подбор гиперпараметров, стэкинг моделей.

Хорошего результата удалось добиться прежде всего за счет огромного количества данных (изначально в объединенном датасете было под миллион строк). Важным моментом было грамотно объединить пять датасетов в один.

Обилие данных позволило удалить излишки с которыми тяжело работать.

Соответственно, точность можно было бы повысить если более скрупулезно вникать в записи. Например поработать со столбцом где дана комплектация автомобиля (очевидно влияющая на цену) или со столбцом где дано от описание владельца (пошариться по ключевым словам).

В общем удалось отработать материал данный в Юните на практике и добиться неплохого результата.