# Отчет по дипломной работе на тему «Предсказание цен на поддержанные автомобили в США в соответствии с данными сайта craigslist.org»

## Введение

Craigslist.org - одна из крупнейших интернет-площадок электронных объявлений в мире, особенно пользующаяся большой популярностью в США. Подавляющую часть этого сайта занимают объявления о продаже поддержанных автомобилей.

Представленный набор данных обновляется каждый месяц и содержит исчерпывающую информацию об автомобилях, которую предоставляет сайт craigslist.org, включая сведения о марке и модели автомобиля, его состоянии, цене продажи, пробеге и т.д. Актуальная версия набора данных представлена от 09.2020.

В рамках выполнения дипломной работы необходимо решить задачу регрессии и построить соответствующую модель, предсказывающую цены на поддержанные автомобили. Для проверки качества регрессии использовать метрики RMSE и MAE.

## Описание проблемы и набор данных

Данная проблема открыта для решения на платформе Kaggle в качестве упражнения для начинающих специалистов в области Data Science. Она была выбрана в качестве темы дипломной работы для подтверждения квалификации в области машинного обучения и умения применения знаний, полученных в ходе обучения.

Набор данных представляет собой 24 признаковых переменных и 1 целевую переменную (конечная стоимость автомобиля), которую необходимо предсказать. Единицы измерения считаются известными и понятными по умолчанию. Список признаковых переменных приведен ниже:

* идентификатор;
* url;
* регион;
* url региона;
* год выпуска;
* производитель;
* модель;
* состояние автомобиля;
* количество цилиндров;
* тип топлива;
* пробег;
* тех.паспорт;
* тип коробки передач;
* VIN-номер;
* привод;
* размер;
* тип кузова;
* цвет;
* url изображения;
* описание;
* округ;
* штат;
* координаты широты;
* координаты долготы.

Всего в наборе 423857 экземпляров. Для построения модели и оценки результатов набор данных был предварительно почищен от выбросов и пропусков, а затем разделен на обучаемую и тестовую выборки. Тестовая выборка составляет 20% от полученного набора данных. В итоговом наборе данных 245795 экземпляров.

## Подготовка данных

Для выполнения поставленной задачи был проведен EDA: выявлены и удалены маловажные признаки и выбросы; построена корреляционная матрица для выявления зависимостей между признаками (Рисунок 1); построены графики распределения числовых переменных (Рисунок 2); построены диаграммы размаха некоторых категориальных переменных и их корреляция с ценой (Рисунок 3, Рисунок 4); построены круговые диаграммы для наблюдения относительных значений каждого признака (Рисунок 5); построен график важности признаков, по сути дублирующий результаты построенной корреляционной матрицы (Рисунок 6).

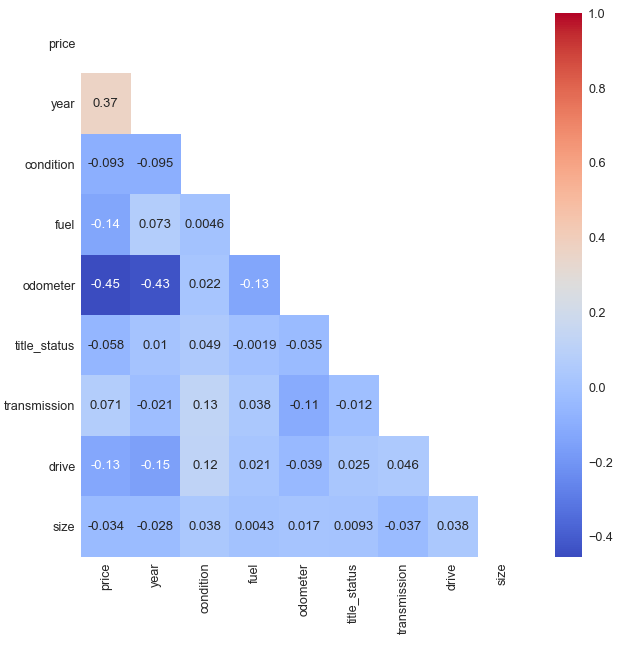


Рисунок 1 – Корреляционная матрица

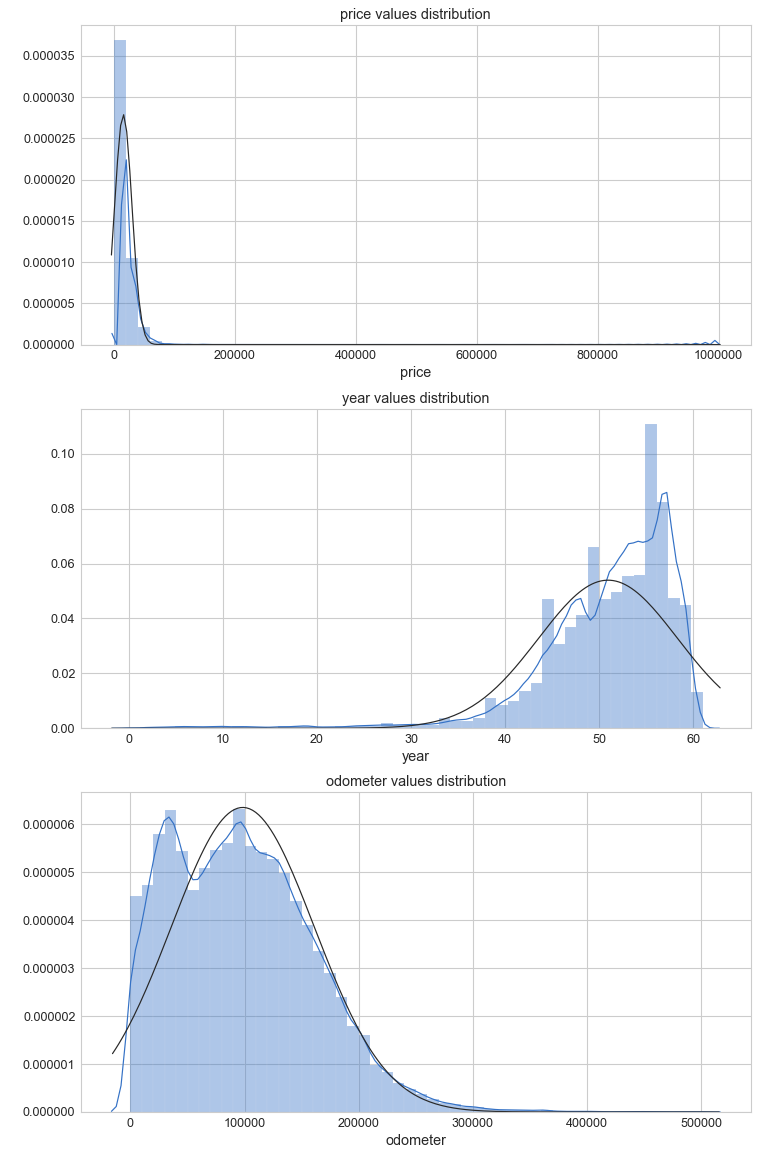


Рисунок 2 – Распределение числовых переменных

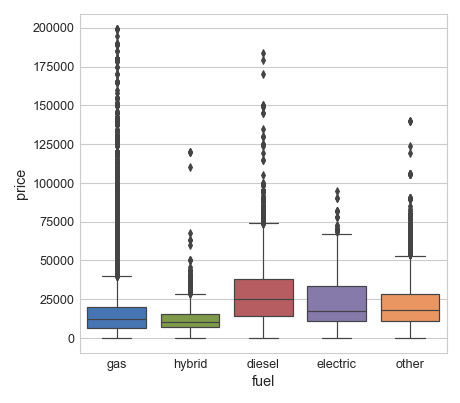


Рисунок 3 – Диаграмма размаха типа топлива

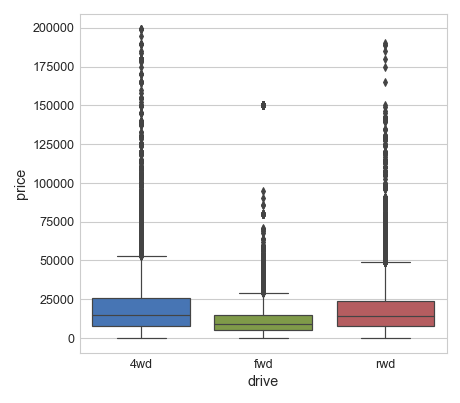


Рисунок 4 – Диаграмма размаха типа привода

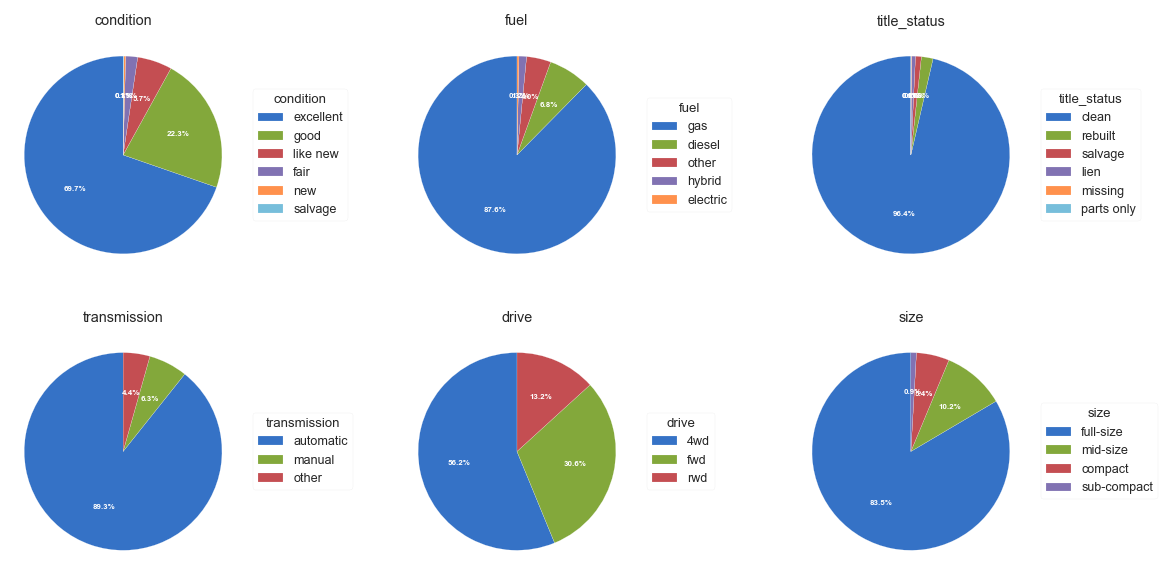


Рисунок 5 – Круговые диаграммы относительных значений признаков

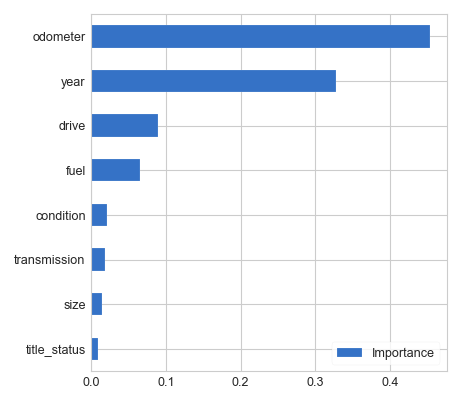


Рисунок 6 – График важности признаков

## Построение модели

Для построения модели и решения задачи регрессии были выбраны линейные алгоритмы Linear Regression и Elastic Net и нелинейный алгоритм Random Forest Regressor, т.к. они являются наиболее подходящими для решения подобного типа задач (в данных отсутствуют выбросы и зашумление, достаточно большое количество экземпляров, небольшое количество значимых переменных).

При построении нелинейной использовались стандартные параметры, за исключением случайного начального состояния.

При построении моделей, основанных на алгоритмах Elastic Net и Random Forest Regressor, были предприняты попытки поиска оптимальных параметров – максимальное количество итераций для Elastic Net и максимальная глубина дерева для Random Forest.

Также построена нелинейная модель, основанная на единственной признаковой переменной – пробеге. При ее построении использовались стандартные параметры.

## Результаты

### Результаты построения моделей

Для оценки результатов модели были выбраны метрики RMSE и MAE – они наиболее эффективно позволяют оценивать результаты предсказаний в случае, когда значения целевой переменной принимают достаточно большие значения, а также явно реагируют на наличие выбросов и шумов.

Полученные значения метрики RMSE приведены на Рисунке 7.



Рисунок 7 – Значения метрики RMSE для линейных и нелинейной моделей

Полученные значения метрики MAE приведены на Рисунке 8.



Рисунок 8 – Значения метрики MAE для линейных и нелинейной моделей

На основе полученных результатов метрик RMSE и MAE можно сделать вывод, что наиболее эффективной является нелинейная модель, построенная с использованием алгоритма Random Forest Regressor.

### Результаты поиска оптимальных параметров

Поиск оптимальных параметров осуществлялся с использованием алгоритма GridSearchCV. Количество стратегий разделения для кросс-валидации равно 10.

Для алгоритма Elastic Net осуществлялся поиск оптимального значения максимального количества итераций в диапазоне от 10 до 1000. В ходе поиска, на который было затрачено 5 мин 58 сек, оптимальным значением максимального количества итераций является значение 10, а полученное значение метрики RMSE равно 11664.3.

Для алгоритма Random Forest Regressor осуществлялся поиск оптимального значения минимальной глубины для дерева в диапазоне от 2 до 30. По графику зависимости значений метрики RMSE от глубины дерева, представленному на Рисунке 9, можно сделать вывод, что результат обучения модели улучшается при увеличении глубины деревьев.

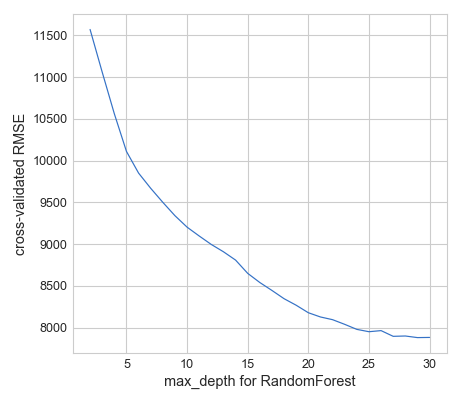


Рисунок 9 – График зависимости RMSE от глубины дерева

В ходе поиска, на который было затрачено 10 ч 1 мин 24 сек, оптимальным значением глубины дерева является значение 29, а полученное значение метрики RMSE равно 7879.3.

### Результаты построения модели с единственным признаком

При оценке результатов построения модели с единственной признаковой переменной – пробегом – также использовались метрики RMSE и MAE. Полученные значения метрик приведены на Рисунке 10.



Рисунок 10 – Значения метрик для построенной модели

На основе полученных результатов можно сделать вывод, что данная модель имеет право на существование, но для достижения лучших результатов необходимо использовать большее количество признаков.

## Дальнейшие направления исследований

Несмотря на то, что выбранные метрики для оценки работы алгоритма показали хорошие результаты, думаю, что, изменяя параметры алгоритма (например, максимальную глубину дерева в совокупности с максимальным количеством листьев), результат может быть улучшен.

Соответственно дальнейшие исследования могут быть направлены на продолжение поиска оптимальных параметров (с использованием алгоритмов GridSearchCV или RandomSearchCV).