

Оценка стоимости автомобиля методом машинного обучения

ИССЛЕДОВАТЕЛЬСКИЕ
РАБОТЫ
УЧАЩИХСЯ

Проект выполнил:
Георгий Шавшишвили,
11-й класс МБОУ «Мурманский политехнический лицей»

Руководитель:
Ермакова Елена Николаевна,
учитель информатики МБОУ МПЛ

Целью проекта является создание прототипа ПО, позволяющего оценить транспортное средство в реальном времени, не требуя подключения к Интернету.

Введение

Данная работа посвящена решению современной задачи построения информационной системы для оценки в режиме реального времени стоимости транспортного средства. Актуальность задачи, с одной стороны, подчёркивается ростом автомобильного рынка, порождающим спрос на быструю и эффективную оценку стоимости транспортных средств, а с другой — отсутствием соответствующих готовых сервисов.

Кому это выгодно?

Клиент: Возможность предсказывать рыночную стоимость подержанных автомобилей может помочь как покупателям, так и продавцам.

Продавцы (дилеры) подержанных автомобилей: Они являются одной из самых больших целевых групп, которые могут быть заинтересованы в результатах этого исследования. Если продавцы подержанных автомобилей лучше понимают, что делает автомобиль желанным, каковы важные особенности подержанного автомобиля, то они могут принять во внимание эти знания и предложить более качественное обслуживание.

Услуги по ценообразованию в Интернете: Существуют веб-сайты, на которых предлагается приблизительная стоимость автомобиля. У них может быть хорошая модель прогнозирования. Однако наличие второй модели может помочь им дать лучший прогноз для своих пользователей. Следовательно, модель, разработанная в этом исследовании, может помочь онлайн-сервисам, которые определяют рыночную стоимость подержанного автомобиля.

Физические лица: Есть много людей, которые интересуются рынком подержанных автомобилей в определённые моменты своей жизни, потому что они хотели про-

дать свою машину или купить подержанную машину. В этом процессе очень важно платить слишком много или продавать меньше рыночной стоимости.

Метод

Обработка датасета, необходимого для обучения ПО; выбор и корректировка математической модели, используемой для получения точных результатов; создание прототипа программного обеспечения с помощью возможностей языка Python, модулей TensorFlow, Scikit-Learn и Matplotlib.

Данные

Данные, используемые в этом проекте, были загружены с Kaggle. CSV файл содержит почти всю необходимую информацию о продаже автомобилей, которую предоставляет Craigslist, включая TTX автомобиля и рекомендованную цену.

Основное содержание

Мы имеем CSV файл с необходимыми данными, но перед использованием его необходимо подготовить — проиндексировать записи об автомобилях, заполним пропуски.

Следующий шаг — разделить данные на наборы для тестов и обучения. Принято следующее распределение данных: 90% — тренировочный набор, 10% — проверка.

Для осуществления машинного обучения будем использовать метод Adaptive Boosting **Алгоритм AdaBoost** (сокр. от adaptive boosting) — алгоритм машинного обучения, предложенный Йоавом Фройндом (Yoav Freund) и Робертом Шапиром (Robert Schapire). Является мета-алгоритмом, в процессе обучения строит композицию из базовых алгоритмов обучения для улучшения их эффективности. AdaBoost является алгоритмом адаптивного бустинга в том смысле, что каждый следующий клас-

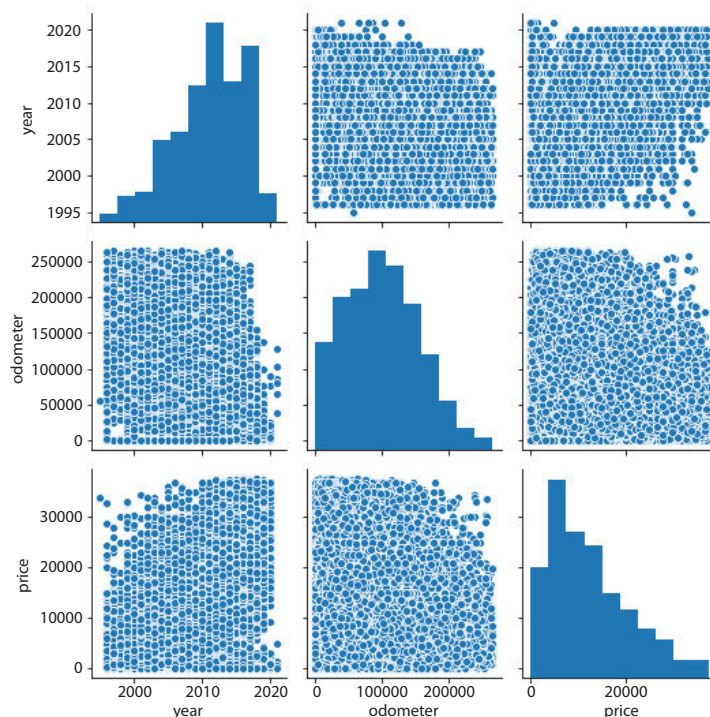


сификатор строится по объектам, которые плохо классифицируются предыдущими классификаторами.

В первую очередь необходимо создать вспомогательную программу, которая наглядно на примере диаграмм покажет нам зависимость цены от технических характеристик автомобиля.

Такая программа была создана и использована. Далее по тексту прилагаются диаграммы построенные с помощью ПО.

Некоторые выводы из набора данных.



Из этого мы ничего не можем сделать. Между переменными нет корреляции.

Из графика распределения мы можем сделать вывод, что сначала цена быстро

растёт, но после определённого момента цена начинает снижаться.

Из диаграммы 1 мы анализируем, что цена на автомобиль с дизельным двигателем выше, чем цена на электрический вариант.

Автомобили с гибридным вариантом имеют самую низкую цену. Из диаграммы 2 мы анализируем, что цена на соответствующее топливо также зависит от состояния автомобиля.

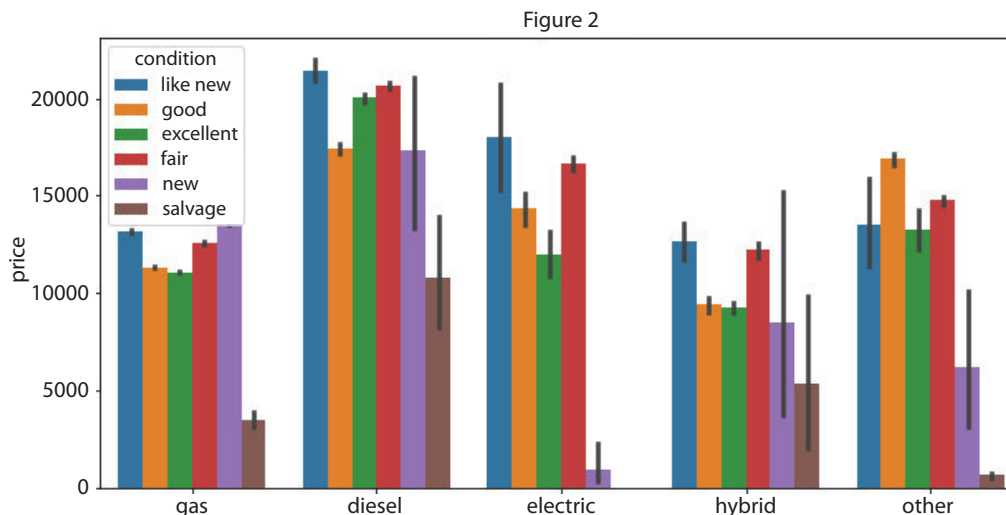
На диаграммы 3 мы анализируем, что цены на автомобили растут в год после 1995 г., а на диаграмме 4 — количество автомобилей также увеличивается в год, и в какой-то момент, например в 2012 г., количество автомобилей почти не изменилось.

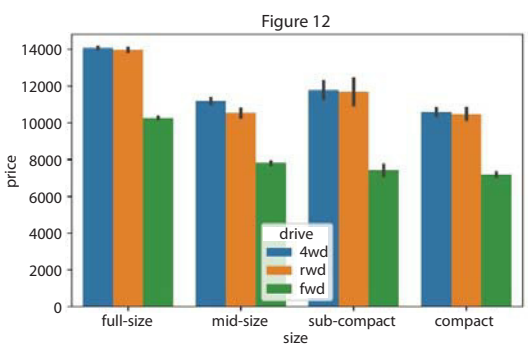
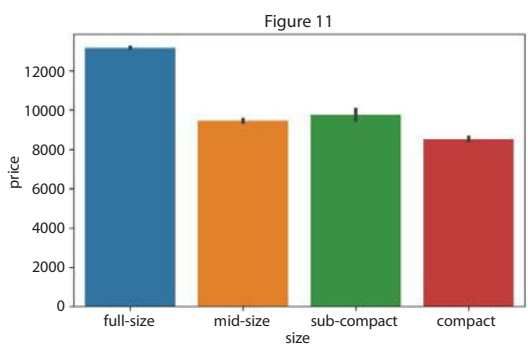
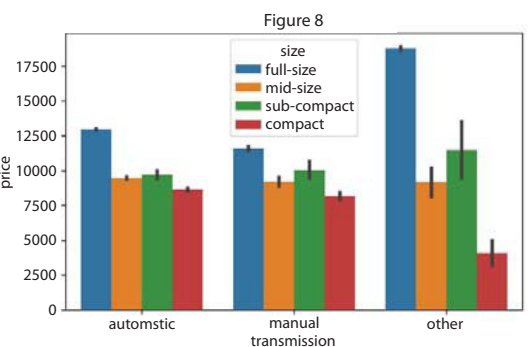
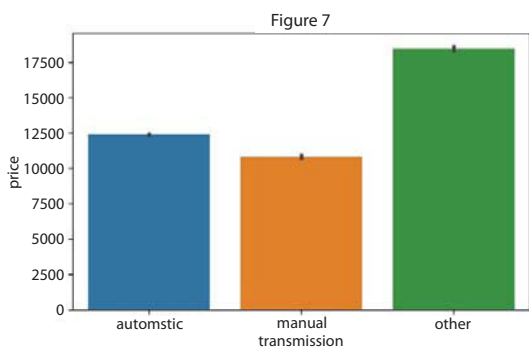
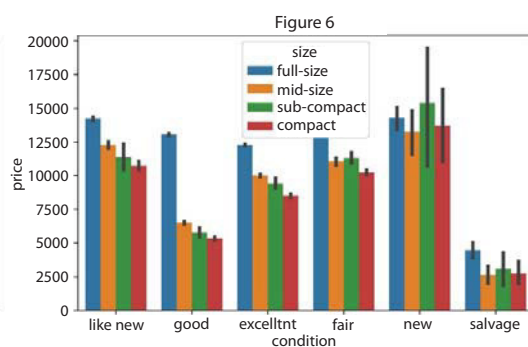
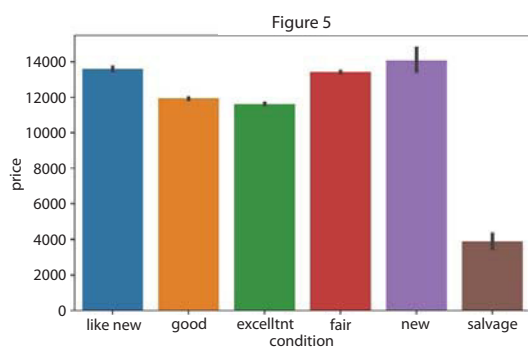
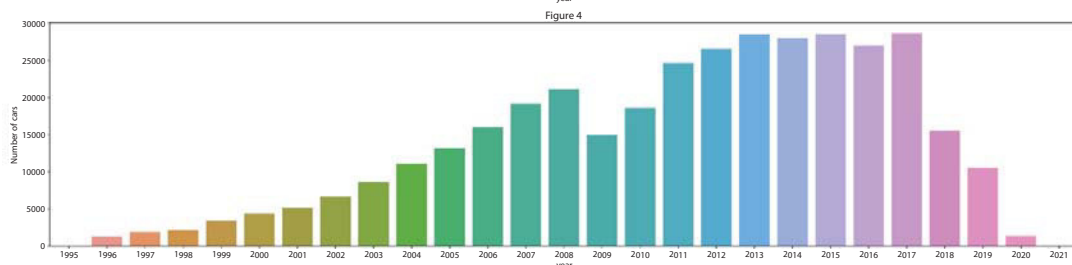
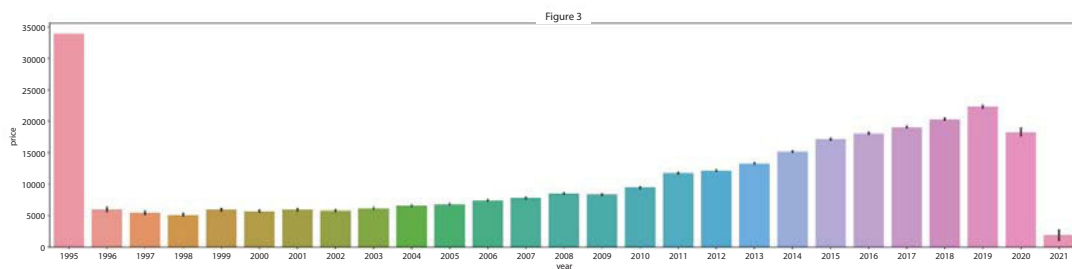
Из диаграммы 5 мы можем проанализировать, что цена автомобилей также зависит от состояния автомобиля, а из рисунка 6 — цена варьируется в зависимости от состояния автомобилей соответствующего размера.

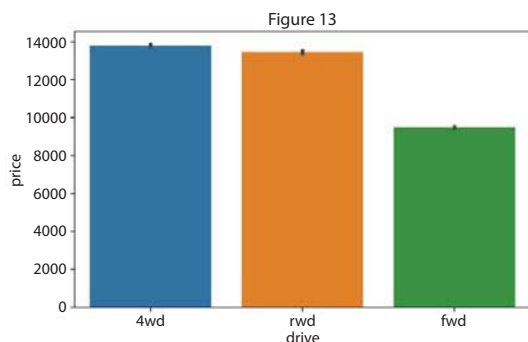
На диаграмме 7–8 мы анализируем, что цена автомобилей также различна для каждой трансмиссии автомобиля. Люди готовы покупать машину с «другой трансмиссией», а цена машин с «механической

трансмиссией» невысока.

Ниже представлены похожие графики с тем же пониманием, но с разными функциями.







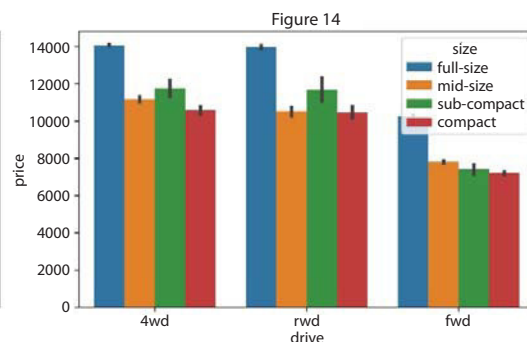
Заключение

Выполняя разные модели машинного обучения, мы стремимся получить лучший результат или меньше ошибок при максимальной точности. Наша цель состояла в том, чтобы спрогнозировать цену подержанных автомобилей, имея 25 предикторов и 509 577 записей данных.

Первоначально выполняется очистка данных для удаления нулевых значений и выбросов из набора данных, затем внедряются модели машинного обучения для прогнозирования цен на автомобили. Затем с помощью визуализации данных были глубоко изучены возможности. Исследуется связь между признаками. 📊

Литература

1. Орельен Ж. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. М.: Альфа-книга, 2018.



Электронные ресурсы:

1. Датасет — Used Cars Dataset; Vehicles listings from Craigslist.org / [Электронный ресурс]. — Режим доступа: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

2. Документация Python / [Электронный ресурс]. — Режим доступа: <https://www.python.org/doc/>

3. Профессиональный вики-ресурс, посвящённый машинному обучению и интеллектуальному анализу данных / [Электронный ресурс]. — Режим доступа: www.Machine-Learning.ru

4. Профессиональный информационно-аналитический ресурс, посвящённый машинному обучению, распознаванию образов и интеллектуальному анализу данных / [Электронный ресурс]. — Режим доступа: <http://www.machinelearning.ru>

5. Ценообразование в автопроме / [Электронный ресурс]. — Режим доступа: <https://www.drive2.ru/b/548910015642075190/>