Итоговая домашняя работа по дисциплине “Эконометрика” на тему:

«Эконометрическое моделирование стоимости подержанных автомобилей»

Работу выполни студент

Дьяков Андрей ПМ22–4

2023 год

Аннотация содержания статьи:

В представленной работе проведено эконометрическое моделирование стоимости подержанных автомобилей с использованием линейной регрессии. Исследование включает в себя анализ влияния различных характеристик, таких как тип кузова, марка и год выпуска, на цену подержанных автомобилей. Методология включает в себя стандартизацию признаков и оценку модели с использованием библиотеки statsmodels. Полученные результаты и оценки производительности модели представлены. Также произведено сравнение влияния признаков на цену автомобиля в зависимости от типа кузова.

Ключевые слова:

Регрессия, автомобиль, параметры, уравнение.

Когда я выбирал тему, я вспомнил, как на сайтах по продаже автомобилей с пробегом я видел рекомендованную цену (например, от 260тыс.рублей до 310тыс.рублей) Меня всегда удивляло, как программа рассчитывает этот диапазон. Поэтому я решил попробовать создать модель предсказания стоимости подержанных автомобилей.

Постановка задачи.

Цель данного исследования – построение математической модели, которая учитывала бы факторы, влияющие на стоимость бывших в употреблении автомобилей различных типов кузова.

Задачи исследования:

- Отобрать факторы для построения модели стоимости б.у. автомобилей.
- Построить многофакторное регрессионное уравнение.
- Оценить модель на адекватность.
- Произвести оценку влияния факторов на стоимость б.у. автомобиля.
- Сделать выводы.

Описание используемых данных:

На сайте kaggle.com была найдена база данных. Сразу же были удалены некоторые столбцы, такие как город (все данные из США), VIN номер и т. п., так как значение города на цену автомобиля не интересует, а VIN номер на цену не влияет. В конечном виде база данных имела следующие колонки:

- Price, цена автомобиля (зависимая переменная y)
- Year, год производства автомобиля x_1
- Manufacturer, фирма производитель x_2
- Condition, состояние автомобиля x_3
- Cylinders, количество цилиндров x_4
- Fuel, тип топлива x_5
- Odometer, пробег x_6
- Transmission, тип трансмиссии x_7
- Drive, привод автомобиля x_8
- Type, тип кузова
- paint color, цвет покраски x_9

Самыми популярными типами кузовов автомобилей на данный момент являются седан и кроссовер. Поэтому из исходной базы данных были созданы две другие, только с седанами и только с кроссоверами.

Расчёт параметров регрессионного уравнения.

Перед расчетом параметров регрессионного уравнения были удалены все строки с пропусками, данные были приведены к численным, год был уменьшен на 1900, так как все автомобили были выпущены после этого года, пробег был поделен нацело на 5000. Выбросы были удалены.

Была построена корреляционные матрицы:

Для седанов:

	price	year	manufacturer	condition	cylinders	fuel	odometer	transmission	drive	paint_color
price	1.000000	0.500703	-0.033427	0.038289	0.327013	-0.001990	-0.263081	0.077958	-0.053667	-0.006618
year	0.500703	1.000000	0.082280	0.127165	-0.176831	0.059497	-0.297308	0.013852	-0.017917	0.020974
manufacturer	-0.033427	0.082280	1.000000	-0.007292	-0.247328	-0.160710	-0.029274	0.065988	-0.076508	0.012243
condition	0.038289	0.127165	-0.007292	1.000000	-0.012074	0.027223	-0.043145	0.047564	0.026056	0.012172
cylinders	0.327013	-0.176831	-0.247328	-0.012074	1.000000	0.050857	0.035219	-0.029979	0.125004	-0.038880
fuel	-0.001990	0.059497	-0.160710	0.027223	0.050857	1.000000	0.015752	-0.008938	-0.004501	0.015487
odometer	-0.263081	-0.297308	-0.029274	-0.043145	0.035219	0.015752	1.000000	-0.017845	0.013715	-0.002425
transmission	0.077958	0.013852	0.065988	0.047564	-0.029979	-0.008938	-0.017845	1.000000	-0.059231	-0.013008
drive	-0.053667	-0.017917	-0.076508	0.026056	0.125004	-0.004501	0.013715	-0.059231	1.000000	0.037666
paint_color	-0.006618	0.020974	0.012243	0.012172	-0.038880	0.015487	-0.002425	-0.013008	0.037666	1.000000

Рисунок 1 Корреляционная матрица седанов

Для кроссоверов:

	price	year	manufacturer	condition	cylinders	fuel	odometer	transmission	drive	paint_color
price	1.000000	0.551829	0.031335	0.035287	0.339255	-0.015740	-0.172749	0.022967	-0.157294	0.021803
year	0.551829	1.000000	0.045474	0.099107	-0.240613	0.044521	-0.231208	-0.022338	0.031492	0.017068
manufacturer	0.031335	0.045474	1.000000	-0.010884	-0.152389	-0.008290	-0.025495	0.037400	-0.117618	0.023195
condition	0.035287	0.099107	-0.010884	1.000000	-0.023030	0.008416	-0.032691	-0.000754	0.048236	0.002107
cylinders	0.339255	-0.240613	-0.152389	-0.023030	1.000000	-0.038341	0.078024	-0.007802	-0.069308	-0.021616
fuel	-0.015740	0.044521	-0.008290	0.008416	-0.038341	1.000000	0.000790	0.008062	0.051705	0.032228
odometer	-0.172749	-0.231208	-0.025495	-0.032691	0.078024	0.000790	1.000000	-0.009731	-0.000799	-0.000185
transmission	0.022967	-0.022338	0.037400	-0.000754	-0.007802	0.008062	-0.009731	1.000000	-0.026803	-0.002296
drive	-0.157294	0.031492	-0.117618	0.048236	-0.069308	0.051705	-0.000799	-0.026803	1.000000	0.035151
paint_color	0.021803	0.017068	0.023195	0.002107	-0.021616	0.032228	-0.000185	-0.002296	0.035151	1.000000

Рисунок 2 Корреляционная матрица кроссоверов

Между независимыми переменными не было выявлено большой корреляции.

Данные были разделены на тестовую и тренировочную выборки, масштабированы. Б ыли построены модели множественных линейных регрессий.

Для кроссоверов:

$$\hat{y} = 16353.3 + 4930.32x_1 + 423.97x_2 - 87.04x_3 + 3795.81x_4 - 144.51x_5 - 696.87x_6 + 257.41x_7 - 988.55x_8 + 177.4x_9 + \varepsilon$$

R–квадрат = 0.575

Для седанов:

$$\hat{y} = 11190.06 + 3197.02x_1 - 172.32x_3 + 2544.05x_4 - 318.69x_5 - 609.83x_6 + 444.22x_7 - 545.95x_8 + \varepsilon$$

R–квадрат = 0.458

Все незначимые независимые переменные были убраны из уравнений, проверены по статистике Стьюдента на пороговом уровне 0.05, уравнения в целом значимы по статистике Фишера. (F табличное = 1.88; F1=1629; F2=1934; значит, уравнения в целом значимы по статистике Фишера)

Анализ адекватности регрессионного уравнения:

Для тестовых данных модель предсказания цены кроссовера показала такой результат:

MSE: 23442614

R–квадрат: 56%

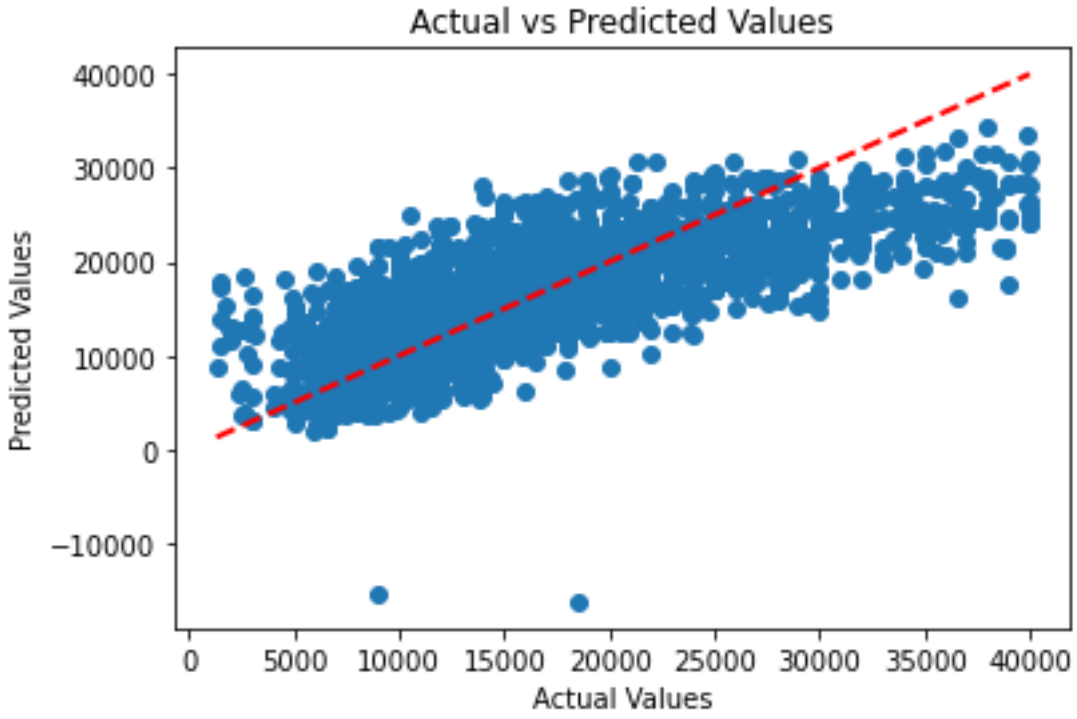

Рисунок 3 предсказанные значения против реальных для модели кроссовера

Для тестовых данных модель предсказания цены седана показала такой результат:

MSE: 18246537

R–квадрат: 47%

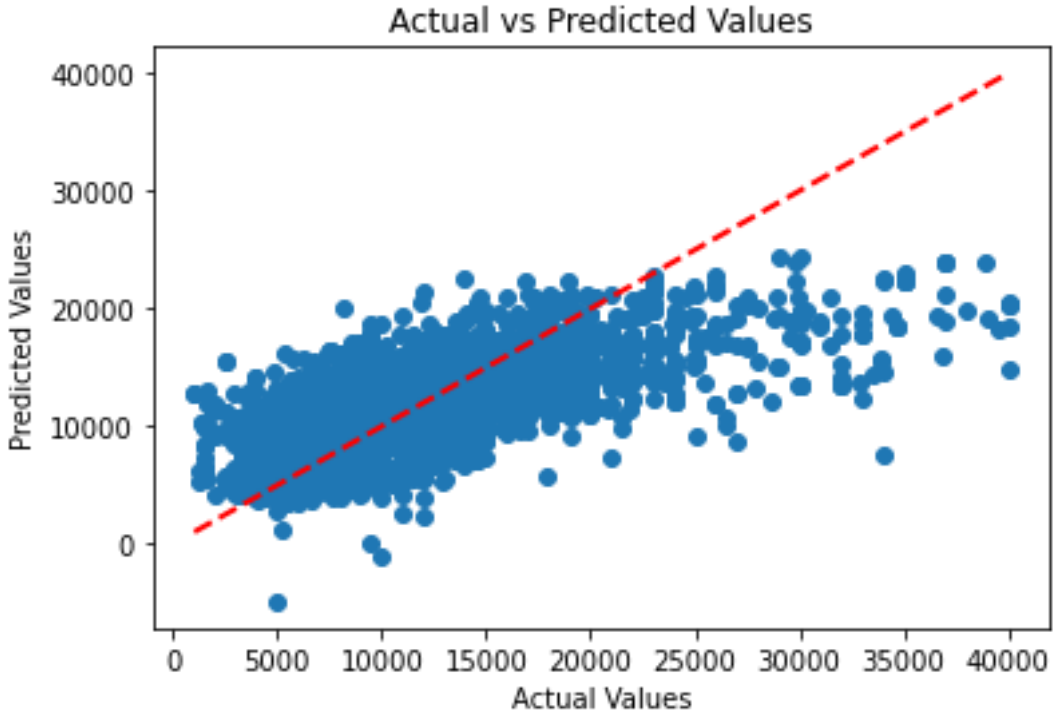

Рисунок 4 предсказанные значения против реальных для модели седана

Значит, обе модели не очень хорошо предсказывают цену подержанного автомобиля.

Проверка условий для получения «хороших» оценок МНК:

проверка на гомоскедастичность

Тест Голдфельда–Квандта для седана:

F–статистика: 0.91

p–значение: 0.999

Тест Голдфельда–Квандта для кроссовера:

F–статистика: 0.99

p–значение: 0.669

Для обоих моделей данные предоставляют недостаточные доказательства в пользу отклонения от гомоскедастичности. Мы не можем отклонить нулевую гипотезу о гомоскедастичности.

Таким образом, на основе теста Голдфельда–Квандта, можно сделать вывод, что в данной модели нет статистически значимых доказательств в пользу гетероскедастичности (непостоянства дисперсии остатков).

Проверка на автокорреляцию:

Статистика теста Дурбина–Ватсона для модели седана: 2

Статистика теста Дурбина–Ватсона для модели кроссовера: 2

Значения близкие к 2 указывают на отсутствие автокорреляции, значит в обоих моделях отсутствует автокорреляция.

Экономический смысл коэффициентов регрессии.

Построение модели множественной линейной регрессии позволяет определить степень влияния каждого фактора на целевую переменную y.

Цена и седанов, и кроссоверов в большей степени зависит от старости автомобиля. Чем новее машина, тем цена выше. Так же большую роль играет количество цилиндров, которые обычно напрямую влияют на кол–во лошадиных сил. Чем цилиндров больше, тем цена выше. Также немаловажным оказался пробег. Чем больше пробег, тем стоимость меньше.

Прогнозирование на основе полученной модели. Доверительный интервал прогноза.

С помощью полученных моделей можно прогнозировать цену автомобиля, но достаточно неточно. Также усложняет прогноз вся предобработка данных.

Расчет доверительных интервал был произведен на python:

Кроссовер: (1761; 30508)

Седан: (–194; 22883)

Выводы.

Рынок подержанных автомобилей представляет собой сложную и динамичную область, где ценообразование зависит от множества факторов. Анализ влияния различных характеристик, таких как марка, тип кузова и год выпуска, на цену подержанных автомобилей позволяет лучше понять динамику этого рынка. В этом исследовании я понял, что цена и седанов, и кроссоверов в большей степени зависит от старости автомобиля. Чем новее машина, тем цена выше. Так же большую роль играет количество цилиндров, которые обычно напрямую влияют на кол–во лошадиных сил. Чем цилиндров больше, тем цена выше. Также немаловажным оказался пробег. Чем больше пробег, тем стоимость меньше. И про тип трансмиссии не стоит забывать.

Построение многофакторного регрессионного уравнения дает инструмент для более точного прогнозирования стоимости подержанных автомобилей. Однако, обнаружено, что не все факторы могут быть учтены, и некоторые важные данные могут оставаться закрытыми или недоступными. Это нам говорит низкий R квадрат.

Важным выводом является необходимость учета не только видимых характеристик, но и факторов, таких как бюджет разработки, рекламные затраты, и другие, которые могут оказывать влияние на формирование цен на подержанные автомобили.

В целом, математическая модель, учитывающая сложные факторы, влияющие на стоимость подержанных автомобилей, может стать полезным инструментом для различных участников рынка, помогая им принимать более информированные решения в этой динамичной сфере.

Источники:

https://cenamashin.ru/statistika/russia/count_cars_kusov

https://stackoverflow.com

https://www.kaggle.com/code/vbmokin/used-cars-price-prediction-by-15-models/notebook

https://fundamental-research.ru/ru/article/view?id=40390

https://auto.ru/mag/article/samym-populyarnym-tipom-kuzova-v-rossii-ostayotsya-sedan

https://www.autostat.ru/news/54414/